EN INTRODUKTION TIL STATISTIK

BIND 2B

Knut Conradsen

4. UDGAVE

LYNGBY 1984

FORELÆSNINGSNOTE

IMSOR

Trykt af [illegible], DTH

INDHOLDSFORTEGNELSE

1	Resumé af den lineære algebra	1.1-1.70
1.1	Vektorrum	1.2
1.1.1	Definition af vektorrum	1.2
1.1.2	Direkte sum af vektorrum	1.6
1.2	Lineære afbildninger og matricer	1.8
1.2.1	Lineære afbildninger	1.8
1.2.2	Matricer	1.10
1.2.3	Lineære afbildningers matrixfremstillinger	1.12
1.2.4	Koordinattransformation	1.13
1.2.5	Rangen af en matrix	1.15
1.2.6	Determinanten af en matrix	1.17
1.2.7	Blokmatricer	1.19
1.3	Pseudoinverse eller generaliseret inverse matrix til en ikke-regulær matrix	1.22
1.4	Egenværdiproblemer. Kvadratiske former	1.35
1.4.1	Egenværdier og egenvektorer for symmetriske matricer	1.35
1.4.2	Singulær-værdi dekomposition af vilkårlig matrix. Q- og R-modus analyser	1.42
1.4.3	Kvadratiske former og positivt definitte matricer	1.46
1.4.4	Det generelle egenværdiproblem for symmetriske matricer	1.54
1.4.5	Sporet af en matrix	1.58
1.4.6	Differentiation af linearform og kvadratisk form	1.59
1.5	Tensor- eller Kronecker produkt af matricer	1.63
1.6	Indre produkter og normer	1.64
2	Flerdimensionale variable	2.1-2.69
2.1	Momenter af flerdimensionale variable	2.1
2.1.1	Middelværdi	2.1
2.1.2	Dispersionsmatricen	2.3
2.1.3	Kovarians	2.6

2.2 Den flerdimensionale normalfordeling 2.10

2.2.1 Definition og simple egenskaber 2.10
2.2.2 Uafhængighed og konturellipsoider 2.18
2.2.3 Betingede fordelinger 2.22
2.2.4 Reproduktivitetssætning og central grænseværdisætning 2.23
2.2.5 Estimation af parametre i en flerdimensional normalfordeling 2.25
2.2.6 Den todimensionale normale fordeling 2.27

2.3 Korrelation og regression 2.34

2.3.1 Den partielle korrelationskoefficient 2.35
2.3.2 Den multiple korrelationskoefficient 2.44
2.3.3 Regression 2.49

2.4 Spaltningssætningen 2.53

2.5 Wishart fordelingen og den generaliserede varians 2.59

2.6 Lidt om estimation af flerdimensionale parametre 2.64

3 Den generelle lineære model 3.1-3.60

3.1 Estimation i den generelle lineære model 3.1

3.1.1 Modelformulering 3.1
3.1.2 Estimation i det regulære tilfælde 3.5
3.1.3 Tilfældet $\underline{x}'\underline{\Sigma}^{-1}\underline{x}$ singulær 3.12
3.1.4 Estimation under bibetingelser 3.25
3.1.5 Konfidensintervaller for forudsagte værdier. Prediktionsinterval 3.32

3.2 Test i den generelle lineære model 3.40

3.2.1 Test for lavere dimension af modelrum 3.40
3.2.2 Successiv testning i den generelle lineære model 3.48

4 Regressionsanalyse 4.1-4.73

4.1 Lineær regressionsanalyse 4.1

4.1.1 Notation og model 4.2
4.1.2 Korrelation og regression 4.6
4.1.3 Analyse af forudsætninger 4.8

4.2 Regression efter ortogonale polynomier 4.14

4.2.1 Definition og modelformulering 4.14
4.2.2 Bestemmelse af ortogonale polynomier 4.19

4.3 Valg af "bedste" regressionsligning 4.27

4.3.1 Problemstillingen 4.27
4.3.2 Undersøgelse af samtlige regressioner 4.30
4.3.3 Backwards elimination 4.32
4.3.4 Forward selection 4.34
4.3.5 Stepwise regression 4.38
4.3.6 Nogle eksisterende programmer 4.41
4.3.7 Numerisk appendix 4.42

4.4 Andre regressionsmodeller og -løsninger 4.50

4.4.1 Ortogonal regression (lineær funktionel relation) 4.50
4.4.2 Ridge-regression 4.56
4.4.3 Ikke-lineær regression og kurvetilpasning 4.66

5 Variansanalyser 5.1-5.83

5.1 Indledning 5.1

5.2 Ensidet variansanalyse 5.3

5.2.1 Modeller 5.3
5.2.2 Analyse af den systematiske model 5.7
5.2.3 Analyse af den tilfældige model 5.11
5.2.4 Resumé af analyserne og et eksempel 5.14

5.3 Tosidet variansanalyse. Hierarkisk klassifikation og krydsklassifikation 5.18

5.3.1 Hierarkisk klassifikation og krydsklassifikation 5.18
5.3.2 Analyse af hierarkisk klassificerede data 5.24
5.3.3 Analyse af krydsklassificerede data 5.31

5.4 Variansanalysemodeller med 3 faktorer 5.39

5.5 Variansanalyser med flere faktorer 5.52

5.5.1 Estimation af parametre og beregning af kvadratafvigelsessummer 5.52
5.5.2 Beregning af forventede værdier af middelkvadratafvigelsessummer 5.58

5.6 Variansanalyseprogrammet ANOVA 5.78

6 Tests i den flerdimensionale normale fordeling 6.1-6.44

6.1 Test for middelværdier 6.1

6.1.1 Hotelling's T^2 i enstikprøvesituationen 6.1
6.1.2 Hotelling's T^2 i tostikprøvesituationen 6.8

6.2 Den flerdimensionale generelle lineære model 6.13

6.3 Variansanalyser for flerdimensionale variable 6.28

6.3.1 Ensidet flerdimensional variansanalyse 6.28
6.3.2 Tosidet flerdimensional variansanalyse 6.30

6.4 Tests vedrørende dispersionsmatricer 6.39

6.4.1 Tests vedrørende en enkelt dispersionsmatrix 6.39
6.4.2 Test for, at flere dispersionsmatricer er ens 6.43

7 Diskriminantanalyse 7.1-7.51

7.1 Diskrimination mellem 2 populationer 7.1

7.1.1 Bayes- og minimaxløsninger 7.1
7.1.2 Diskrimination mellem 2 normale populationer 7.4
7.1.3 Diskrimination med ukendte parametre 7.15
7.1.4 Test for bedste diskriminatorfunktion 7.19
7.1.5 Test for yderligere information 7.22

7.2 Diskrimination mellem flere populationer 7.24

7.2.1 Bayesløsning 7.24
7.2.2 Bayesløsning i tilfældet med flere normale fordelinger 7.26
7.2.3 Alternativ diskriminationsprocedure i tilfældet med flere populationer 7.32

7.3 Nogle standardprogrammer til beregning af lineære diskriminatorer 7.36

8 Principale komponenter, kanoniske variable og korrelationer samt faktoranalyse 8.1-8.64

8.1 Principale komponenter 8.2

8.1.1 Definition og simple egenskaber 8.2
8.1.2 Estimation og testning 8.7

8.2 Kanoniske variable og korrelationer 8.17

8.3 Faktoranalyse 8.20

8.3.1 Model og forudsætninger 8.21
8.3.2 Estimation af faktorer (faktorvægte) 8.24
8.3.3 Faktor rotation 8.27
8.3.4 Beregning af faktorværdier (factor scores) 8.33
8.3.5 Et case-study 8.39
8.3.6 Lidt om maximum likelihood faktoranalyse 8.48
8.3.7 Q-modus analyse 8.51
8.3.8 Nogle standardprogrammer 8.54

9 Statistisk analyse af tidsrækker 9.1-9.179

9.1 Fourier-transformationen, forskydningsoperatorer og lineære systemer 9.2

9.1.1 Fourier-transformationen 9.2
9.1.2 Forskydningsoperatorer 9.14
9.1.3 Tidsvarianter, lineære systemer 9.19
9.1.4 Sampling problemet 9.24

9.2 Stokastiske processer og deres momentfunktioner 9.29

9.2.1 Kort om tidsrækker og stokastiske processer 9.29
9.2.2 Den diskrete lineære proces. AR-, MA- og ARMA-processer 9.44
9.2.3 Den kontinuerte lineære proces 9.59
9.2.4 Estimation af autokovariansfunktionen 9.63

9.3 Den klassiske analyse 9.67

9.3.1 Udjævning og trend 9.67
9.3.2 Den klassiske dekomposition 9.80

9.4	Endimensional spektralanalyse	9.88
9.4.1	Spektret for en stokastisk proces	9.88
9.4.2	Estimation af (power-)spektre	9.100
9.5	Filtrering	9.116
9.6	Krydsspektralanalyse	9.132
9.6.1	Krydskovarians og krydskorrelation	9.132
9.6.2	Estimation af krydskovariansfunktion	9.136
9.6.3	Krydsspektret	9.137
9.6.4	Estimation af krydsspektret	9.141
9.6.5	Eksempel på krydsspektralanalyse	9.143
9.7	Box-Jenkins' metode	9.152
9.7.1	ARIMA-processer	9.152
9.7.2	Sæsonmodellen	9.156
9.7.3	Forudsigelser i ARIMA-processer	9.158
9.7.4	Identifikation af og estimation i en ARIMA-proces	9.164
9.7.5	Et eksempel	9.168

Index

KAPITEL 6

Tests i den flerdimensionale normale fordeling

I dette kapitel skal vi give en række generaliseringer af nogle velkendte teststørrelser baseret på endimensionale, normalt fordelte stokastiske variable. I de fleste tilfælde vil teststørrelserne være umiddelbare analogier til de velkendte, blot skal multiplikation erstattes med matrixmultiplikation, numerisk værdi med determinant af matrix etc.

6.1 Test for middelværdier

6.1.1 Hotelling's T^2 i enstikprøvesituationen

I dette afsnit skal vi betragte uafhængige stokastiske variable $\underline{X}_1, \cdots, \underline{X}_n$, hvor

$$\underline{X}_i \in N_p(\underline{\mu}, \underline{\underline{\Sigma}}) ,$$

d.v.s. p-dimensionalt normalt fordelt med middelværdivektor $\underline{\mu}$ og dispersionsmatrix $\underline{\underline{\Sigma}}$. Det forudsættes, at $\underline{\underline{\Sigma}}$ er regulær og ukendt. Vi ønsker at teste en hypotese om, at middelværdivektoren $\underline{\mu}$ er lig en given vektor $\underline{\mu}_0$ mod alle alternativer, i.e.

$$H_0 : \underline{\mu} = \underline{\mu}_0 \quad \text{mod} \quad H_1 : \underline{\mu} \neq \underline{\mu}_0$$

Vi repeterer først nogle resultater om estimatorerne. Fra sætning 2.27 p. 2.61 har vi følgende resultater om den empiriske middelværdivektor $\underline{\bar{X}}$ og den empiriske dispersionsmatrix $\underline{\underline{S}}$

$$\underline{\bar{X}} = \frac{1}{n} \sum_{i=1}^{n} \underline{X}_i \qquad \in N_p(\underline{\mu}, \frac{1}{n} \underline{\underline{\Sigma}})$$

$$\underline{\underline{S}} = \frac{1}{n-1} \sum_{i=1}^{n} (\underline{X}_i - \underline{\bar{X}})(\underline{X}_i - \underline{\bar{X}})' \in W(n-1, \frac{1}{n-1} \underline{\underline{\Sigma}})$$

$\underline{\bar{X}}$ og $\underline{\underline{S}}$ er stokastisk afhængige.

I det følgende har vi endvidere brug for følgende resultat om fordelingen af visse funktioner af normalt fordelte og Wishart-fordelte stokastiske variable

Lemma 6.1 Lad $\underline{Y}$ være en p-dimensional stokastisk variabel og lad U være en p×p stokastisk matrix med

$$\underline{Y} \in N_p(\underline{\mu}, \underline{\underline{\Sigma}})$$

$$m\underline{\underline{U}} \in W(m, \underline{\underline{\Sigma}}) ,$$

og lad endvidere $\underline{Y}$ og $\underline{\underline{U}}$ være stokastisk uafhængige. Vi sætter

$$T^2 = \underline{Y}' \, \underline{\underline{U}}^{-1} \, \underline{Y} .$$

Da gælder

$$\frac{m-p+1}{mp} T^2 \in F(p, m-p+1 \; ; \; \underline{\mu}' \underline{\underline{\Sigma}}^{-1} \underline{\mu}) ,$$

d.v.s. venstresiden er ikke-centralt F-fordelt med skævhedsparameteren $\underline{\mu}' \underline{\underline{\Sigma}}^{-1} \underline{\mu}$ og frihedsgrader (p, m-p+1). Hvis $\underline{\mu} = \underline{0}$, er skævhedsparameteren 0, d.v.s. vi har da specielt

$$\frac{m-p+1}{mp} T^2 \in F(p, m-p+1) .$$

Bevis Forbigås. Se e.g. Andersson (1958) p. 106.

Vi har nu følgende hovedresultat

Sætning 6.2 Vi anvender betegnelsen

$$T^2 = n(\underline{\bar{X}} - \underline{\mu}_0)' \underline{\underline{S}}^{-1} (\underline{\bar{X}} - \underline{\mu}_0) ,$$

hvor $\underline{\bar{X}}$, $\underline{\mu}_0$ og $\underline{\underline{S}}$ er som anført i indledningen til dette afsnit. Da er det kritiske område for kvotienttestet af H_0 med H_1 på niveau α lig

$$C = \{\underline{x}_1, \cdots, \underline{x}_n \mid \frac{n-p}{(n-1)p} t^2 > F(p,n-p)_{1-\alpha}\} ,$$

hvor t^2 er den observerede værdi af T^2.

Bevis Af lemma 6.1 følger, at

$$\frac{n-p}{(n-1)p} T^2 \in F(p,n-p)$$

under H_0. Heraf følger, at C er kritisk område for et test af H_0 mod H_1 på niveau α. At det svarer til kvotienttestet følger ved direkte regning bl.a. under benyttelse af sætning 1.2.

Q.E.D.

Bemærkning 1 Størrelsen T^2 kaldes ofte Hotelling's T^2 efter Harold Hotelling, der først betragtede denne teststørrelse.

Bemærkning 2 I det endimensionale tilfælde anvender vi teststørrelsen

$$Z = \frac{\sqrt{n}\ (\bar{X}-\mu_0)}{S} .$$

Vi har nu, at Z^2 kan skrives

$$Z^2 = n(\bar{X}-\mu_0)[S^2]^{-1}(\bar{X}-\mu_0) \ ,$$

d.v.s. præcis det samme som T^2 reducerer til i det endimensionale tilfælde. Bemærk endvidere, at kvadratet på en student-fordelt variabel $t(\nu)$ er $F(1,\nu)$-fordelt, hvorfor der (selvfølgelig) også er overensstemmelse mellem teststørrelsernes fordelinger.

Af hensyn til beregninger af teststørrelsen kan det være nyttigt at erindre sig følgende sætning, hvoraf det fremgår, at inversion af en matrix kan "erstattes" af beregning af nogle determinanter.

Sætning 6.3 Lad betegnelserne være som ovenfor. Da gælder

$$T^2 = \frac{\det[\underline{\underline{S}}+n(\underline{\bar{X}}-\underline{\mu}_0)(\underline{\bar{X}}-\underline{\mu}_0)']}{\det[\underline{\underline{S}}]} - 1$$

Bevis Forbigås. Rent teknisk og følger ved anvendelse af sætning 1.2 p.1.21 på matricen

$$\begin{bmatrix} -1 & \sqrt{n}(\underline{\bar{X}}-\underline{\mu}_0)' \\ \sqrt{n}(\underline{\bar{X}}-\underline{\mu}_0) & \underline{\underline{S}} \end{bmatrix}$$

Q.E.D.

Vi anfører nu et illustrativt

Eksempel 6.1 I nedenstående tabel er anført værdier for silicium- og aluminiumindholdet (i %) i 7 stikprøver indsamlet på månen

	Prøve						
	1	2	3	4	5	6	7
Silicium	19.4	21.5	19.2	18.4	20.6	19.8	18.7
Aluminium	5.9	4.0	4.0	5.4	6.2	5.7	6.0

Det er nu af stor interesse at erfare, om disse prøver kan antages at stamme fra en population med samme middelværdier, som gælder for jordisk basalt. Disse er

$$\underline{\mu}_0 = \begin{pmatrix} 22.10 \\ 7.40 \end{pmatrix} .$$

Det synes rimeligt at anvende Hotelling's T^2 til at afgøre ovenstående spørgsmål. Kaldes observationerne $\underline{x}$, ··· , $\underline{x}$, finder vi

$$\bar{\underline{x}} = \begin{pmatrix} 19.66 \\ 5.31 \end{pmatrix} ,$$

$$\underline{\underline{s}} = \begin{pmatrix} 1.1795 & -0.3076 \\ -0.3076 & 0.8681 \end{pmatrix} .$$

Da

$$\bar{\underline{x}} - \underline{\mu}_0 = \begin{pmatrix} -2.44 \\ -2.09 \end{pmatrix} ,$$

er

$$n(\bar{\underline{x}}-\underline{\mu}_0)(\bar{\underline{x}}-\underline{\mu}_0)' = \begin{pmatrix} 41.68 & 35.70 \\ 35.70 & 30.58 \end{pmatrix} ,$$

og

$$\underline{\underline{s}} + n(\bar{\underline{x}}-\underline{\mu}_0)(\bar{\underline{x}}-\underline{\mu}_0)' = \begin{pmatrix} 42.86 & 35.39 \\ 35.39 & 31.45 \end{pmatrix} .$$

Følgelig er

$$t^2 = \frac{95.49}{0.9293} - 1 = 101.75 .$$

F-teststørrelsen er

$$\frac{7-2}{6 \cdot 2} t^2 = 42.8 > F(2,5)_{.999} = 37.1 ,$$

og hypotesen forkastes derfor i det mindste på alle niveauer α større end 0.1%. Det forekommer derfor ikke rimeligt at antage, at de 7 måneprøver stammer fra en population med samme middelindhold af silicium og aluminium som jordisk basalt.

□

Ud fra resultatet i sætning 6.2 konstrueres let konfidensområder for $\underline{\mu}$. Vi har med den sædvanlige notation

Sætning 6.4 Et $(1-\alpha)$-konfidensområde for forventningen $E(\underline{X})$ er

$$\{\underline{\mu} \mid n(\bar{\underline{x}}-\underline{\mu})' \underline{\underline{S}}^{-1} (\bar{\underline{x}}-\underline{\mu}) \leq \frac{(n-1)p}{n-p} F(p,n-p)_{1-\alpha}\} ,$$

d.v.s. en ellipsoide med centrum i $\bar{\underline{x}}$ og med hovedakser bestemt af egenvektorerne i den inverse empiriske dispersionsmatrix.

Bevis Triviel følge af definitionen på et konfidensområde og sætning 6.2.

Q.E.D.

Vi fortsætter nu eksempel 6.1 i nedenstående

Eksempel 6.2 Vi vil nu bestemme et 95% konfidensområde for middelværdivektoren. Ifølge sætning 6.4 er konfidensområdet begrænset af ellipsen

$$7(19.66-\mu_1,\ 5.31-\mu_2)\underline{\underline{S}}^{-1}\begin{pmatrix} 19.66-\mu_1 \\ 5.31-\mu_2 \end{pmatrix} = \frac{12}{5} F(2,5)_{.95}$$

eller

$$(19.66-\mu_1,\ 5.31-\mu_2)\underline{\underline{S}}^{-1}\begin{pmatrix} 19.66-\mu_1 \\ 5.31-\mu_2 \end{pmatrix} = 1.9851 .$$

Vi finder

$$\underline{\underline{S}}^{-1} = \begin{pmatrix} 0.9341 & 0.3310 \\ 0.3310 & 1.2692 \end{pmatrix}$$

med egenværdierne 1.4727 og 0.7307 og tilsvarende (normerede) egenvektorer

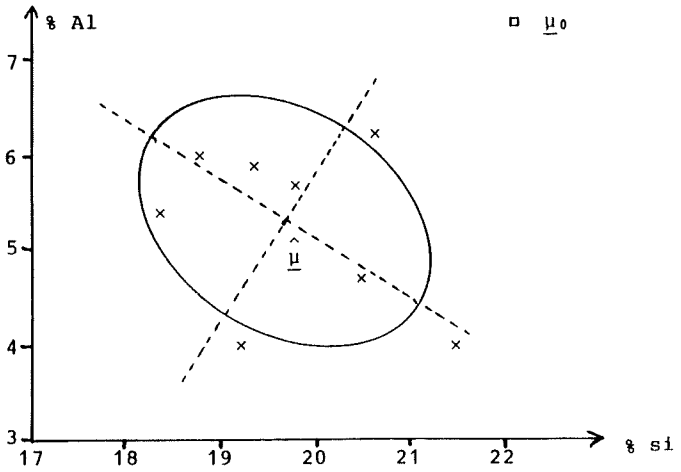
$$\begin{pmatrix} 0.5236 \\ 0.8520 \end{pmatrix} \quad \text{og} \quad \begin{pmatrix} -0.8520 \\ 0.5236 \end{pmatrix} .$$

I koordinatsystemet med origo i $\underline{\bar{x}}$ og med ovenstående vektorer som enhedsvektorer har ellipsen ligningen

$$1.4727\ y_2^2 + 0.7307\ y_2^2 = 1.9851$$

eller

$$\frac{y_1^2}{1.1610^2} + \frac{y_2^2}{1.6482^2} = 1$$



I ovenstående tegning er konfidensområdet og observationerne anført. Desuden er $\underline{\mu}_0 = (22.10,\ 7.40)'$ anført. Det ses, at dette punkt ligger uden for konfidensområdet i overensstemmelse med,

at hypotesen $\underline{\mu} = \underline{\mu}_0$ mod $\underline{\mu} \neq \underline{\mu}_0$ forkastedes på alle niveauer større end 0.01% og dermed specielt for $\alpha = 5\%$.

□

6.1.2 Hotelling's T^2 i tostikprøvesituationen

Ganske analogt til t-testet i det endimensionale tilfælde kan Hotelling's T^2 også anvendes til at undersøge, om stikprøver fra to normale populationer (med samme dispersionsstruktur) kan antages at have samme forventningsværdier.

Vi betragter indbyrdes uafhængige stokastiske variable $\underline{X}_1, \cdots, \underline{X}_n$ og $\underline{Y}_1, \cdots, \underline{Y}_m$, hvor

$$\underline{X}_i \in N_p(\underline{\mu}, \underline{\underline{\Sigma}})$$

$$\underline{Y}_i \in N_p(\underline{\mu}, \underline{\underline{\Sigma}}) ,$$

og vi ønsker at teste

$$H_0 : \underline{\mu} = \underline{\nu} \quad \text{mod} \quad H_1 : \underline{\mu} \neq \underline{\nu} .$$

Vi anvender betegnelserne

$$\overline{\underline{X}} = \frac{1}{n} \sum_{i=1}^{n} \underline{X}_i$$

$$\overline{\underline{Y}} = \frac{1}{m} \sum_{i=1}^{m} \underline{Y}_i$$

$$\underline{\underline{S}}_1 = \frac{1}{n-1} \sum_{i=1}^{n} (\underline{X}_i - \overline{\underline{X}})(\underline{X}_i - \overline{\underline{X}})'$$

$$\underline{\underline{S}}_2 = \frac{1}{m-1} \sum_{i=1}^{m} (\underline{Y}_i - \overline{\underline{Y}})(\underline{Y}_i - \overline{\underline{Y}})'$$

$$\underline{\underline{S}} = \frac{(n-1)\underline{\underline{S}}_1 + (m-1)\underline{\underline{S}}_2}{n+m-2}$$

Ifølge sætning 2.27 og sætning 2.26 har vi

$$\underline{\bar{X}} \in N_p(\underline{\mu}, \frac{1}{n}\underline{\underline{\Sigma}})$$

$$\underline{\bar{Y}} \in N_p(\underline{\mu}, \frac{1}{m}\underline{\underline{\Sigma}})$$

$$\underline{\underline{S}} \in W(n+m-2, \frac{1}{n+m-2}\underline{\underline{\Sigma}}) \ .$$

Vi formulerer nu hovedresultatet om testning af H_0 mod H_1 i

Sætning 6.5 Vi anvender de samme betegnelser som giver ovenfor. Vi sætter

$$T^2 = \frac{nm}{n+m}(\underline{\bar{X}}-\underline{\bar{Y}})' \underline{\underline{S}}^{-1}(\underline{\bar{X}}-\underline{\bar{Y}}) \ .$$

Da er det kritiske område for test af H_0 mod H_1 på niveau α lig

$$C = \{\underline{x}_1, \cdots, \underline{x}_n, \underline{y}_1, \cdots, \underline{y}_m \mid \frac{n+m-p-1}{(n+m-2)p} t^2 > F(p,n+m-p-1)_{1-\alpha}\}$$

Her er t^2 den observerede værdi af T^2.

Bevis Af lemma 6.1 og ovenfor anførte relationer følger, at

$$\frac{n+m-p-1}{(n+m-2)p} T^2 \in F(p,n+m-p-1; (\underline{\mu}-\underline{\nu})' \underline{\underline{\Sigma}}^{-1}(\underline{\mu}-\underline{\nu})) \ ,$$

og heraf følger resultatet umiddelbart.

Q.E.D.

Ganske som i enstikprøvesituationen kan vi også her benytte resultatet til at bestemme et konfidensområde for differensen mellem middelværdivektorerne. Vi har nemlig

Sætning 6.6 Vi betragter fremdeles den ovenfor anførte situation og sætter $\underline{\mu}-\underline{\nu} = \underline{\delta}_o$. Da er et $(1-\alpha)$-konfidensområde for $\underline{\delta}_o$ lig

$$\{\underline{\delta} \mid \frac{nm}{n+m}\,(\bar{\underline{x}}-\bar{\underline{y}}-\underline{\delta})'\underline{\underline{s}}^{-1}(\bar{\underline{x}}-\bar{\underline{y}}-\underline{\delta}) \leq \frac{(n+m-2)p}{n+m-p-1}\,F(p,n+m-p-1)_{1-\alpha}\} \ .$$

Bevis Direkte følge af definitionen på et konfidensområde og sætning 6.5.

Q.E.D.

Bemærkning 1 Konfidensområdet er en ellipsoide med centrum i $\bar{\underline{x}}-\bar{\underline{y}}$ og hovedakser bestemt af egenvektorerne i $\underline{\underline{s}}^{-1}$.

Bemærkning 2 De anførte testresultater og konfidensintervaller kræver som anført, at dispersionsmatricerne for $\underline{X}$- og for $\underline{Y}$-observationerne er ens. Hvis dette ikke er tilfældet, er ovenstående resultater ikke eksakte, og en anden fremgangsmåde må anvendes. Dette skal vi ikke komme ind på her, men vil blot henvise e.g. til Andersson (1958), p. 118.

Vi vil nu betragte et eksempel på anvendelsen af T^2 i en tostikprøvesituation.

Eksempel 6.3 På Laboratoriet for Varme- og Klimateknik, DtH, har man ved et klimaforsøg målt følgende

(i) højden i cm,

(ii) fordampningstab i g/m^2hud i 3 timer,

(iii) middeltemperatur i ^oC. Denne temperatur fås ved at måle hudtemperaturen 14 forskellige steder hvert 5'te minut (samme steder hver gang). Middeltemperaturen er således et gennemsnit af i alt 14×5 = 70 målinger,

på 16 mænd og 16 kvinder. Resultatet af forsøget er givet i tabellen p. 6.11.

Person nr.	Højde i cm	Fordampningstab i g/m^2hud	Middeltemperatur i ^oC
1	177	18.1	33.9
2	189	18.8	33.2
3	181	20.4	33.9
4	184	19.5	33.8
5	183	30.5	33.3
6	178	22.2	33.6
7	162	19.4	34.2
8	176	26.7	33.2
9	190	16.6	33.2
10	180	45.4	33.5
11	179	24.0	33.9
12	175	34.6	33.8
13	183	21.3	33.5
14	177	33.3	33.9
15	185	22.9	33.8
16	176	18.6	33.5
1	160	14.6	32.9
2	171	27.0	33.7
3	168	27.6	32.3
4	171	20.2	33.1
5	169	30.8	33.4
6	169	17.4	33.5
7	167	21.1	33.0
8	170	19.3	34.1
9	162	21.5	33.8
10	160	15.2	33.0
11	168	15.4	33.7
12	157	25.2	33.9
13	161	13.9	34.8
14	164	20.2	31.9
15	161	25.3	34.0
16	180	12.6	33.5

Vi opfatter disse tal som realisationer af stokastiske variable

$$\underline{X}_1, \cdots, \underline{X}_{16} \quad \text{og} \quad \underline{Y}_1, \cdots, \underline{Y}_{16} .$$

Vi antager ydermere, at de variable er stokastisk uafhængige, og at

$$\underline{X}_i \in N(\underline{\mu}, \underline{\underline{\Sigma}})$$

og

$$\underline{Y}_i \in N(\underline{\nu}, \underline{\underline{\Sigma}}) \ ,$$

d.v.s. at dispersionsmatricerne antages at være ens. Vi skal senere diskutere rimeligheden af denne hypotese.

Estimaterne for $\underline{\mu}$ og $\underline{\nu}$ er de empiriske middelværdier, d.v.s.

$$\hat{\underline{\mu}} = \overline{\underline{x}} = \begin{pmatrix} 179.7 \\ 24.5 \\ 33.6 \end{pmatrix}$$

og

$$\hat{\underline{\nu}} = \overline{\underline{y}} = \begin{pmatrix} 166.1 \\ 20.5 \\ 33.4 \end{pmatrix} \ .$$

Vi vil nu undersøge, om forskellen mellem $\hat{\underline{\mu}}$ og $\hat{\underline{\nu}}$ er signifikant, i.e. om det kan antages, at $\underline{\mu}$ og $\underline{\nu}$ er ens.

Vi finder med de i sætning 6.5 valgte betegnelser

$$\underline{\underline{s}} = \begin{pmatrix} 38.5 & -4.3 & -0.8 \\ -4.3 & 45.5 & -0.3 \\ -0.8 & -0.3 & 0.3 \end{pmatrix} \ ,$$

og dermed

$$t^2 = \frac{16 \cdot 16}{16+16} \, (\overline{\underline{x}} - \overline{\underline{y}})' \underline{\underline{s}}^{-1} (\overline{\underline{x}} - \overline{\underline{y}}) = 52.4 \ .$$

Teststørrelsen bliver

$$\frac{16+16-3-1}{(16+16-2)\,3} \, 52.4 = 16.3 \ .$$

Da

$$F(3,28)_{0.999} = 7.19$$

vil en hypotese om, at $\underline{\mu} = \underline{\nu}$ i det mindste blive forkastet på alle niveauer større end 0.1%. Vi vil derfor konkludere, at der er væsentlig forskel på de 3 variable for mænd og for kvinder, et resultat, der næppe chokerer nogen, når det erindres, at den første variabel angiver højden.

Betragtes i stedet kun 2'den og 3'die koordinaterne, d.v.s. værdierne for fordampningstab og middeltemperatur, fås test-størrelsen

$$\frac{16+16-2-1}{(16+16-2)2}\ (4.0,\ 0.2)\begin{pmatrix} 45.5 & -0.3 \\ -0.3 & 0.3 \end{pmatrix}^{-1}\begin{pmatrix} 4.0 \\ 0.2 \end{pmatrix} \simeq 0.2\ .$$

Denne størrelse skal sammenlignes med fraktilerne i en F(2,29)-fordeling, og det ses straks, at en hypotese om, at middelværdi-vektorerne er ens, accepteres på alle rimelige niveauer.

□

6.2 Den flerdimensionale generelle lineære model

I de foregående afsnit har vi betragtet en- og tostikprøvesi-tuationen for den flerdimensionale normale fordeling, og vi har set, at de flerdimensionale resultater er helt analoge til de endimensionale. I dette og det følgende afsnit skal vi fortsætte denne analogi og udlede resultater vedrørende re-gressions- og variansanalyser af flerdimensionale variable.

Vi betragter indbyrdes uafhængige observationer $\underline{Y}_1, \cdots, \underline{Y}_n$,

$$\underline{Y}_i \in N_p(\underline{\mu}_i, \underline{\underline{\Sigma}})\ .$$

Dispersionsmatricen $\underline{\underline{\Sigma}}$ (og middelværdivektorerne $\underline{\mu}_i$) antages ukendte. Vi ordner observationerne i en n×p datamatrix

$$\underline{\underline{Y}} = \begin{bmatrix} \underline{Y}_1' \\ \vdots \\ \underline{Y}_n' \end{bmatrix} = \begin{bmatrix} Y_{11} & \cdots & Y_{p1} \\ \vdots & & \vdots \\ Y_{1n} & \cdots & Y_{pn} \end{bmatrix} .$$

Her repræsenterer de enkelte rækker altså fx. gentagelser af målinger af et p-dimensionalt fænomen. I fuld analogi med den model, der er betragtet under den endimensionale generelle lineære model, antager vi, at middelværdiparametrene $\underline{\mu}_i$ kan skrives som kendte lineære funktioner af andre (og færre) ukendte parametre $\underline{\underline{\theta}}$, d.v.s.

$$E(\underline{\underline{Y}}) = \underline{\underline{x}}\ \underline{\underline{\theta}} = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \theta_{11} & \cdots & \theta_{1p} \\ \vdots & & \vdots \\ \theta_{k1} & \cdots & \theta_{kp} \end{bmatrix} .$$

Her forudsættes altså $\underline{\underline{x}}$ kendt og $\underline{\underline{\theta}}$ ukendt. Denne model kan anskues fra flere vinkler. Sætter vi den j'te søjle i $\underline{\underline{Y}}$-matricen lig

$$\underline{\underline{Y}}_{j|} = \begin{bmatrix} Y_{j1} \\ \vdots \\ Y_{jn} \end{bmatrix} ,$$

kan vi skrive

$$E(\underline{\underline{Y}}_{j|}) = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \theta_{1j} \\ \vdots \\ \theta_{kj} \end{bmatrix} = \underline{\underline{x}}\ \underline{\underline{\theta}}_{j|} .$$

De n målinger på den j'te "egenskab" vil derfor følge en almindelig endimensional generel lineær model.

Skriver vi i stedet op middelværdien for den enkelte observation $\underline{Y}_i$, finder vi

$$E(\underline{Y}_i') = \begin{pmatrix} x_{i1} & \cdots & x_{ik} \end{pmatrix} \begin{pmatrix} \theta_{11} & \cdots & \theta_{1p} \\ \vdots & & \vdots \\ \theta_{k1} & \cdots & \theta_{kp} \end{pmatrix} = \underline{x}_i' \underline{\underline{\theta}} ,$$

hvor $\underline{x}_i' = \underline{\underline{x}}_{-i}$ er den i'te række i $\underline{\underline{x}}$-matricen. Dette giver umiddelbart

$$E(\underline{Y}_i) = \underline{\underline{\theta}}' \underline{x}_i ,$$

hvilket er en analog til den endimensionale regressionsmodel.

Ordnes observationerne i en søjlevektor

$$\underset{\sim}{Y} = vc(\underline{\underline{Y}}) = \begin{bmatrix} \underline{\underline{Y}}_{1|} \\ \vdots \\ \underline{\underline{Y}}_{p|} \end{bmatrix} ,$$

får vi af sætning 2.7, p. 2.9, at

$$D(\underset{\sim}{Y}) = \underline{\underline{\Sigma}} \otimes \underline{\underline{I}}_n ,$$

hvor $\underline{\underline{\Sigma}} \otimes \underline{\underline{I}}_n$ er tensorproduktet af $\underline{\underline{\Sigma}}$ og $\underline{\underline{I}}_n$, jvf. afsnit 1.5.

Et første problem er at estimere $\underline{\underline{\theta}}$. Der gælder

Sætning 6.7 Vi betragter ovenstående situation. Hvis observationerne $\underline{Y}_i$ er normalt fordelte, er maximum likelihood skønnet for $\underline{\underline{\theta}}$ givet ved

$$\hat{\underline{\underline{\theta}}} = (\underline{\underline{x}}'\underline{\underline{x}})^{-1}\underline{\underline{x}}'\underline{\underline{Y}} .$$

Bevis Forbigås. Se fx. Anderson (1958).

Bemærkning 1 Vi ser, at

$$\hat{\underline{\underline{\theta}}}_{j|} = (\underline{\underline{x}}'\underline{\underline{x}})^{-1}\underline{\underline{x}}'\underline{\underline{Y}}_{j|} ,$$

d.v.s. estimatet for den j'te søjle i $\underline{\underline{\theta}}$ er simpelt hen lig det resultat, vi får ved kun at betragte den endimensionale generelle lineære model for den j'te "egenskab".

Bemærkning 2 Hvis observationerne ikke er normalt fordelte, vil man stadig kunne bruge det anførte skøn $\hat{\underline{\theta}}$, idet denne selvfølgelig ligesom i det endimensionale tilfælde besidder en Gauss-Markov egenskab. Vi skal ikke gå i detaljer med dette, men dog nævne et par resultater. Mindste kvadraters egenskaber bliver, at

$$M = (\underline{Y} - \underline{\underline{X}}\underline{\theta})'(\underline{Y} - \underline{\underline{X}}\underline{\theta}) - (\underline{Y} - \underline{\underline{X}}\hat{\underline{\theta}})'(\underline{Y} - \underline{\underline{X}}\hat{\underline{\theta}})$$

er positiv semidefinit. Dette medfører, at

$$ch_i(\underline{Y} - \underline{\underline{X}}\underline{\theta})'(\underline{Y} - \underline{\underline{X}}\underline{\theta}) \geq ch_i(\underline{Y} - \underline{\underline{X}}\hat{\underline{\theta}})(\underline{Y} - \underline{\underline{X}}\hat{\underline{\theta}}) \ ,$$

hvor ch_i betegner den i'te største egenværdi. Dette medfører igen, at $\hat{\underline{\theta}}$ minimaliserer

$$\det(\underline{Y} - \underline{\underline{X}}\underline{\theta})'(\underline{Y} - \underline{\underline{X}}\underline{\theta})$$

og

$$tr(\underline{Y} - \underline{\underline{X}}\underline{\theta})'(\underline{Y} - \underline{\underline{X}}\underline{\theta}) \ .$$

Bemærkning 3 Vi har ovenfor stiltiende forudsat, at $\underline{\underline{X}}'\underline{\underline{X}}$ har fuld rang, d.v.s. at $rg(\underline{\underline{X}}) = k < n$. Hvis dette ikke er tilfældet, kan man i analogi med de endimensionale resultater anføre løsninger ved hjælp af pseudoinverse matricer.

Efter disse betragtninger over estimation af $\hat{\underline{\theta}}$ vender vi os mod estimation af $\underline{\underline{\Sigma}}$.

Sætning 6.8 Vi betragter situationen fra sætning 6.7. Da er maximum likelihood skønnet for $\underline{\underline{\Sigma}}$ lig

$$\hat{\underline{\underline{\Sigma}}}^* = \frac{1}{n}\sum_{i=1}^{n}(\underline{Y}_i - \hat{\underline{\theta}}'\underline{x}_i)(\underline{Y}_i - \hat{\underline{\theta}}'\underline{x}_i)'$$

$$= \frac{1}{n}(\underline{Y} - \underline{\underline{X}}\hat{\underline{\theta}})'(\underline{Y} - \underline{\underline{X}}\hat{\underline{\theta}})$$

$$= \frac{1}{n}[\underline{Y}'\underline{Y} - (\underline{\underline{X}}\hat{\underline{\theta}})'(\underline{\underline{X}}\hat{\underline{\theta}})] \ .$$

Det (i,j)'te element kan også skrives

$$\hat{\sigma}^*_{ij} = \frac{1}{n}(\underline{Y}_{|i} - \underline{x}\hat{\underline{\theta}}_{|i})'(\underline{Y}_{|j} - \underline{x}\hat{\underline{\theta}}_{|j}) \ .$$

Bevis De mange identiteter mellem $\hat{\underline{\Sigma}}$'s elementer fremgår ved simple matrixmanipulationer. For selve resultatet henvises til Anderson (1958).

Fordelingen af de anførte estimatorer anføres i

Sætning 6.9 Vi betragter situationen fra sætningerne 6.7 og 6.8, og vi indfører de sædvanlige betegnelser

$$\underset{\sim}{\theta} = vc(\underline{\theta}) = \begin{bmatrix} \underline{\theta}_{|1} \\ \vdots \\ \underline{\theta}_{|p} \end{bmatrix}$$

$$\hat{\underset{\sim}{\theta}} = vc(\hat{\underline{\theta}}) = \begin{bmatrix} \hat{\underline{\theta}}_{|1} \\ \vdots \\ \hat{\underline{\theta}}_{|p} \end{bmatrix} \ .$$

Da gælder, at $\hat{\underset{\sim}{\theta}}$ er normalt fordelt

$$\hat{\underset{\sim}{\theta}} = vc(\hat{\underline{\theta}}) \in N_{pk}(\underset{\sim}{\theta}, \underline{\Sigma} \otimes (\underline{x}'\underline{x})^{-1}) \ ,$$

og $n\hat{\underline{\Sigma}}^*$ er Wishart fordelt

$$n\hat{\underline{\Sigma}}^* \in W(\underline{\Sigma}, n-k) \ .$$

Endelig er $\underline{\Sigma}^*$ og $\hat{\underset{\sim}{\theta}}$ og dermed $\underline{\Sigma}^*$ og $\hat{\underline{\theta}}$ stokastisk uafhængige.

Bevis Det er trivielt, at

$$E(\hat{\underline{\theta}}) = E[(\underline{x}'\underline{x})^{-1}\underline{x}'\underline{Y}] = (\underline{x}'\underline{x})^{-1}\underline{x}'\underline{x}\underline{\theta} = \underline{\theta}$$

medfører, at $E(\hat{\underset{\sim}{\theta}}) = \underset{\sim}{\theta}$. Endvidere er selvsagt $\hat{\underset{\sim}{\theta}}$ normalt fordelt.

Ydermere gælder

$$D(\hat{\underline{\theta}}_{|j}) = \sigma_{ii}(\underline{x}'\underline{x})^{-1}$$

og

$$C(\hat{\underline{\theta}}_{|i}, \hat{\underline{\theta}}_{|j}) = (\underline{x}'\underline{x})^{-1}\underline{x}'C(\underline{Y}_{|i}, \underline{Y}_{|j})\underline{x}(\underline{x}'\underline{x})^{-1} = \sigma_{ij}(\underline{x}'\underline{x})^{-1} .$$

Heraf fås resultatet vedrørende dispersionsmatricen for $\hat{\underset{\sim}{\theta}}$ umiddelbart.

Resultatet vedrørende fordelingen af $\hat{\underline{\Sigma}}^*$ og vedrørende uafhængigheden af $\hat{\underline{\theta}}$ og $\hat{\underline{\Sigma}}^*$ er helt analoge til de tilsvarende endimensionale resultater, men vi vil ikke komme nærmere ind på dette her. Læseren henvises til fx. Anderson (1958).

Q.E.D.

Af sætningen fås umiddelbart

Corollar Det centrale skøn for $\underline{\Sigma}$ er lig

$$\hat{\underline{\Sigma}} = \frac{n}{n-k}\hat{\underline{\Sigma}}^* = \frac{1}{n-k}(\underline{Y} - \underline{x}\underline{\theta})'(\underline{Y} - \underline{x}\underline{\theta}) .$$

Bevis Følger trivielt, når det erindres, at

$$E(W(k, \underline{\Delta})) = k\underline{\Delta} .$$

Q.E.D.

Vi vender os dernæst mod testning af parametrene i modellen.

Der gælder

Sætning 6.10 Vi betragter den foranstående situation inklusive forudsætningen om observationernes normalitet. Endvidere betragtes hypotesen

$$H_0 : \underline{A}\,\underline{\theta}\,\underline{B}' = \underline{C} \quad \text{mod} \quad H_1 : \underline{A}\,\underline{\theta}\,\underline{B}' \neq \underline{C} ,$$

hvor $\underline{\underline{A}}(r\times k)$, $\underline{\underline{B}}(s\times p)$ og $\underline{\underline{C}}(r\times s)$ er givne matricer. Vi indfører betegnelserne

$$\underline{\underline{\Delta}} = \underline{\underline{A}}\ \hat{\underline{\underline{\theta}}}\ \underline{\underline{B}}' - \underline{\underline{C}}$$

$$\underline{\underline{R}} = n\hat{\underline{\underline{\Sigma}}}^* = (\underline{Y} - \underline{\underline{x}}\underline{\underline{\theta}})'(\underline{Y} - \underline{\underline{x}}\underline{\underline{\theta}})$$

og

$$\underline{\underline{S}} = \underline{\underline{B}}\ \underline{\underline{R}}\ \underline{\underline{B}}'$$

$$\underline{\underline{H}} = \underline{\underline{\Delta}}'[\underline{\underline{A}}(\underline{\underline{x}}'\underline{\underline{x}})^{-1}\underline{\underline{A}}']^{-1}\underline{\underline{\Delta}}\ .$$

Da er kvotienttestet for test af H_0 mod H_1 ækvivalent med testet givet ved det kritiske område

$$\{\underline{y} \mid \frac{\det(\underline{\underline{s}})}{\det(\underline{\underline{s}}+\underline{\underline{h}})} \leq U(s,r,n-p)_\alpha\}\ ,$$

hvor $U(s,r,n-p)_\alpha$ er α-fraktilen i nulhypotesefordelingen af teststørrelsen (se nedenfor).

Bevis Forbigås. Det bygger essentielt på, at det kan vises, at $\underline{\underline{S}}$ og $\underline{\underline{H}}$ er uafhængige Wishart fordelte variable, hvis H_0 er sand. For nærmere detaljer må henvises til litteraturen.

Som det indirekte fremgår af sætningens formulering, er nulhypotesefordelingen af

$$U = \frac{\det(\underline{\underline{S}})}{\det(\underline{\underline{S}}+\underline{\underline{H}})}$$

alene afhængig af s, r og n-p. Størrelsen benævnes i litteraturen Wilk's Λ eller Anderson's U. Da fordelingen indeholder 3 parametre, er den noget besværlig at arbejde med i praksis, og vi anfører derfor en approximation med en F-fordeling i nedenstående

Sætning 6.11 Lad U være U(p,q,r)-fordelt og sæt

$$t = \begin{cases} 1 , & p^2 + q^2 = 5 \\ \sqrt{\frac{p^2q^2-4}{p^2+q^2-5}} , & p^2 + q^2 \neq 5 \end{cases} .$$

$$v = \tfrac{1}{2}(2r + q - p - 1) .$$

Da er

$$F = \frac{1 - U^{1/t}}{U^{1/t}} \cdot \frac{vt + 1 - \frac{1}{2}pq}{pq}$$

approximativt fordelt som

$$F(pq,\ vt + 1 - \tfrac{1}{2}pq) .$$

Hvis enten p eller q er lig 1 eller 2, er approximationen eksakt.

Bevis Forbigås.

Vi skal nu illustrere de indførte betreber i nedenstående eksempel.

Eksempel 6.4 I perioden 1968-69 blev der ved Landbohøjskolens forsøgsgård for planteavl, Højbakkegård, gennemført et vækstforsøg med lucerne. I alt undersøgte man afkom efter 176 krydsninger, og for at fastlægge "kvaliteten" af de enkelte krydsninger måltes 9 egenskaber på hver af disse. De 9 variable er anført i nedenstående tabel.

For de 5 første variables vedkommende er der som anført tale om en "karaktergivning". Denne fremgangsmåde er valgt, da det er svært at måle de pågældende variable direkte, og erfaringsmæssigt giver denne fremgangsmåde tilfredsstillende resultater.

Variabel nr. & navn	Måleenhed	Forklaring
1: Væksttype	Karakter 1-9	1 = nedliggende vækst, 9 = opretstående vækst
2: Genvækst efter vinter	"	1 = dårligst, 9 = bedst
3: Krybeevne	"	1 = ingen udløbere, 9 = flest udløbere
4: Vigør	"	1 = svagest, 2 = stærkest
5: Blomstringstid	"	1 = seneste blomstring, 2 = tidligste blomstring
6: Plantehøjde	cm	
7: Frøvægt	g pr. plante	
8: Plantevægt	g pr. plante efter tørring	
9: Procent frø	%	Beregnet for hver plante ved hjælp af (7) og (8)

De følgende analyser er baseret på gennemsnitsværdier for de 9 variable baseret på tal fra mellem 15 og 20 planter (de fleste resultater er baseret på 20 planter). I nedenstående tabel er anført et udsnit af disse tal.

Obs.nr = krydsnings-nr.	Variabel nr. og navn								
	1 Vækst-type	2 Gen-vækst	3 Kry-be-evne	4 Vigør	5 Blom-string	6 Plan-te-højde	7 Frø-vægt	8 Plan-te-vægt	9 Pro-cent frø
1	4.11	5.00	3.05	6.17	3.67	50.00	3.47	120.10	2.75
2	3.08	4.75	4.17	7.50	5.17	61.50	0.82	111.33	0.75
3	3.12	4.00	3.35	6.53	3.94	55.29	0.86	97.47	0.81
.									
.									
.									
176	4.00	4.40	4.60	7.40	2.90	50.00	0.66	153.50	0.44

Et overordnet formål med forsøget har været at undersøge samvariationen mellem de 9 variable. Mere specifikt har man bl.a. været interesseret i, hvorledes variabel 3 (Krybeevne) og variabel 4 (Vigør) varierer sammen med de øvrige. De to anførte variable er som regel af stor betydning for en plantes almindelige udvikling, og det er derfor væsentligt, hvorledes sammenhængen er med de øvrige variable.

Som en første orientering bestemmer vi den empiriske korrelationsmatrix. Den er fundet til

	1	2	3	4	5	6	7	8	9
1	1.000	-0.033	0.116	0.018	0.131	-0.207	0.035	-0.087	0.041
2	-0.033	1.000	0.711	0.515	0.125	0.199	-0.025	0.348	-0.066
3	0.116	0.711	1.000	0.440	0.022	0.039	-0.133	0.218	-0.157
4	0.018	0.515	0.440	1.000	0.201	0.517	0.071	0.689	-0.081
5	0.131	0.125	0.022	0.201	1.000	0.496	0.487	0.168	0.486
6	-0.207	0.199	0.039	0.517	0.496	1.000	0.453	0.559	0.367
7	0.035	-0.025	-0.133	0.071	0.487	0.453	1.000	0.360	0.947
8	-0.087	0.348	0.218	0.689	0.168	0.559	0.360	1.000	0.128
9	0.041	-0.066	-0.157	-0.081	0.486	0.367	0.947	0.128	1.000

Vi ser, at variabel 1 (vækstype) kun er svagt korreleret med de øvrige variable, hvorimod fx. variabel 2 og 3 (genvækst og krybeevne) samt (naturligvis) 7 og 9 (frøvægt og % frø) er stærkt korrelerede.

Vi er som nævnt specielt interesserede i variabel 3's og variabel 4's samvariation med de øvrige variable. Vi bemærker, at der er en række halvstore korrelationer, men det er svært at danne sig et indtryk alene på basis af disse. Vi vil derfor forsøge, om det er muligt at udtrykke disse to variable som lineære funktioner af de øvrige, d.v.s.

$$E(Y_1) = \sum_{i=1}^{n} \theta_{i1} x_i$$

$$E(Y_2) = \sum_{i=1}^{n} \theta_{i2} x_i \ ,$$

hvor vi nu har anvendt variabelbetegnelserne

Y_1 = krybeevne

Y_2 = vigør

x_1 = vækstype

x_2 = genvækst efter vinter

x_3 = blomstringstid

x_4 = plantehøjde

x_5 = frøvægt

x_6 = plantevægt

x_7 = procent frø .

Der er her åbenbart tale om en flerdimensional generel lineær model. Sætter vi $\underline{\underline{\theta}} = (\theta_{ij})$, fås

$$\hat{\underline{\underline{\theta}}} = \begin{bmatrix} 0.28400 & 0.42731 \\ 0.79508 & 0.22230 \\ -0.02573 & 0.02607 \\ -0.01151 & 0.06290 \\ -0.14467 & -0.16756 \\ 0.00307 & 0.01103 \\ 0.10614 & 0.03463 \end{bmatrix} .$$

Antages

$$\begin{pmatrix} Y_{1i} \\ Y_{2i} \end{pmatrix} \in N(\underline{\mu}_i, \underline{\underline{\Sigma}}) \quad ,$$

bliver det centrale skøn over $\underline{\underline{\Sigma}}$ lig

$$\hat{\underline{\underline{\Sigma}}} = \begin{bmatrix} 0.85897 & 0.07870 \\ 0.07870 & 0.29444 \end{bmatrix} .$$

Matricen $(\underline{\underline{x}}'\underline{\underline{x}})^{-1}$ er fundet til

1	2	3	4	5	6	7
1.55920	-0.16549	-0.47258	-0.05010	0.41826	-0.00235	-0.42289
-0.16549	0.85139	-0.17981	-0.01327	0.63774	-0.01759	-0.69467
-0.47258	-0.17981	1.77862	-0.10728	-0.29340	0.01164	-0.02184
-0.05010	-0.01327	-0.10728	0.02253	0.12325	-0.00441	-0.17012
0.41826	0.63774	-0.29340	0.12325	5.25546	-0.08437	-7.04885
-0.00235	-0.01759	0.01164	-0.00441	-0.08437	0.00243	0.11182
-0.42289	-0.69467	-0.02184	-0.17012	-7.04885	0.11182	10.11541

Herved kan let beregnes varians og kovarians på de enkelte θ-værdier. Vi har jo

$$D(\underset{\sim}{\hat{\theta}}) = \underline{\underline{\Sigma}} \otimes (\underline{\underline{X}}'\underline{\underline{X}})^{-1} = \begin{pmatrix} \sigma_{11}(\underline{\underline{X}}'\underline{\underline{X}})^{-1} & \sigma_{12}(\underline{\underline{X}}'\underline{\underline{X}})^{-1} \\ \sigma_{21}(\underline{\underline{X}}'\underline{\underline{X}})^{-1} & \sigma_{22}(\underline{\underline{X}}'\underline{\underline{X}})^{-1} \end{pmatrix} ,$$

og dermed fx.

$$\hat{V}(\hat{\theta}_{42}) = 0.2944 \cdot 0.02253 = 0.0066 .$$

Disse resultater kan bruges ved konstruktion af almindelige t-tests for de enkelte koefficienter. Dette skal vi dog ikke komme ind på her. Vi vil i stedet give et par eksempler på, hvorledes man konstruerer simultane tests. Lad os e.g. betragte hypotesen

$$H_0 : \theta_{41} = \theta_{42} = 0$$

mod alle alternativer. Den skal bringes på den i sætning 6.10 angivne form. Vi får dette ved at vælge

$$\underline{\underline{A}} = (0,0,0,1,0,0,0)$$

$$\underline{\underline{B}} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

og

$$\underline{\underline{C}} = (0,0) \quad .$$

Da bliver nemlig

$$\underline{\underline{A}}\ \underline{\underline{\theta}}\ \underline{\underline{B}}' = (\theta_{41},\theta_{42}) \quad .$$

Ved anvendelse af et standardprogram (BMDX63) fås F-teststørrelsen

$$F = 53.66$$

med frihedsgrader

$$(f_1,f_2) = (2,168) \quad .$$

Teststørrelsen er her eksakt F-fordelt, da $s = 2$ og $r = 1$. Det ses, at den observerede F-værdi er signifikant på alle rimelige niveauer.

Som et andet eksempel betragtes hypotesen

$$\underline{\underline{\theta}}_1 = \begin{bmatrix} \theta_{51} & \theta_{52} \\ \theta_{61} & \theta_{62} \\ \theta_{71} & \theta_{72} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

mod alle alternativer. Denne hypotese kan bringes på den i sætning 6.10 anførte form ved at vælge

$$\underline{\underline{A}} = \begin{bmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} ,$$

$$\underline{\underline{B}} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

og

$$\underline{\underline{C}} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} ;$$

thi da er

$$\underline{\underline{A}}\ \underline{\underline{\theta}}\ \underline{\underline{B}}' = \underline{\underline{\theta}}_1 \ .$$

Med det før omtalte standardprogram findes

$$F = 10.63 \ ; \quad (f_1, f_2) = (6,336) \ .$$

Der er således fremdeles klar signifikans.

Som et sidste eksempel betragtes hypotesen

$$\theta_{62} = \theta_{72} = 0$$

mod alle alternativer. Denne bringes på standardformen ved at vælge

$$\underline{\underline{A}} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} ,$$

$$\underline{\underline{B}} = \begin{pmatrix} 0 & 1 \end{pmatrix}$$

og

$$\underline{\underline{C}} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} .$$

F-teststørrelsen har (2,169) frihedsgrader og er fundet til 27.4. De anførte værdier er derfor signifikante.

□

6.3 Variansanalyser for flerdimensionale variable

Vi skal nu specialisere de i det foregående afsnit fundne resultater til generaliseringer af de endimensionale en- og tosidede variansanalyser. Først

6.3.1 Ensidet flerdimensional variansanalyse

Vi betragter observationer

$$\begin{matrix} \underline{Y}_{11} , \cdots , \underline{Y}_{1n_1} \\ \vdots \qquad\qquad \vdots \\ \underline{Y}_{k1} , \cdots , \underline{Y}_{kn_k} \end{matrix} .$$

Disse antages at være stokastisk uafhængige med

$$\underline{Y}_{ij} \in N_p(\underline{\mu}_i, \underline{\underline{\Sigma}}) , \quad i = 1,\cdots,k ; \quad j = 1,\cdots,n_i ,$$

d.v.s. p-dimensionalt normalt fordelte med samme dispersionsmatrix.

Vi ønsker at teste hypotesen

$$H_0 : \underline{\mu}_1 = \cdots = \underline{\mu}_k$$

mod

$$H_1 : \exists i,j(\underline{\mu}_i \neq \underline{\mu}_j) .$$

Vi definerer helt analogt til ensidede variansanalyser "kvadratafvigelsesmatricerne"

$$\underline{\underline{T}} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\underline{Y}_{ij} - \overline{\underline{Y}})(\underline{Y}_{ij} - \overline{\underline{Y}})'$$

$$\underline{\underline{W}} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\underline{Y}_{ij} - \overline{\underline{Y}}_i)(\underline{Y}_{ij} - \overline{\underline{Y}}_i)'$$

$$\underline{\underline{B}} = \sum_{i=1}^{k} n_i (\overline{\underline{Y}}_i - \overline{\underline{Y}})(\overline{\underline{Y}}_i - \overline{\underline{Y}})' \quad .$$

Her er - med $n = \sum_i n_i$ -

$$\overline{\underline{Y}}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \underline{Y}_{ij}$$

$$\overline{\underline{Y}} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} \underline{Y}_{ij} \quad .$$

Ved lidt regning ses, at "total"-matricen $\underline{\underline{T}}$ er summen af "imellem grupper" matricen $\underline{\underline{B}}$ og "inden for grupper" matricen $\underline{\underline{W}}$, i.e.

$$\underline{\underline{T}} = \underline{\underline{W}} + \underline{\underline{B}} \quad ,$$

d.v.s. vi har som i det endimensionale tilfælde en spaltning af den totale variation i variationen mellem grupper og variationen inden for grupper.

Det er trivielt, at vi som centralt skøn over dispersionsmatricen $\underline{\underline{\Sigma}}$ kan anvende

$$\hat{\underline{\underline{\Sigma}}} = \frac{1}{n-k} \underline{\underline{W}} \quad .$$

Hvis hypotesen er sand, vil også $\underline{\underline{T}}$ være proportional med et sådant skøn. Hvis hypotesen ikke er sand, vil $\underline{\underline{T}}$ være "større". Derfor forekommer følgende sætning vel nok intuitivt rimelig

Sætning 6.12 Kvotienttestet for test af hypotesen H_0 mod H_1 er givet ved det kritiske område

$$\{\underline{Y}_{11},\cdots,\underline{Y}_{kn_k} \mid \frac{\det(\underline{\underline{W}})}{\det(\underline{\underline{T}})} \leq U(p,k-1,n-k)_\alpha\} \quad .$$

Bevis Forbigås. Følger ved valg af særlige $\underline{\underline{A}}$, $\underline{\underline{B}}$ og $\underline{\underline{C}}$ matricer i sætning 6.10.

Som for endimensionale variansanalyser samles resultaterne i et Variansanalyseskema

Variationskilde	SAK-matrix	Frihedsgrader
Afvigelse fra hypotesen = variation mellem grupper	$\underline{\underline{B}} = \sum_i n_i (\underline{\bar{Y}}_i - \underline{\bar{Y}})(\underline{\bar{Y}}_i - \underline{\bar{Y}})'$	k-1
Fejl = variation inden for grupper	$\underline{\underline{W}} = \sum_i \sum_j (\underline{Y}_{ij} - \underline{\bar{Y}}_i)(\underline{Y}_{ij} - \underline{\bar{Y}}_i)'$	n-k
Total	$\underline{\underline{T}} = \sum_i \sum_j (\underline{Y}_{ij} - \underline{\bar{Y}})(\underline{Y}_{ij} - \underline{\bar{Y}})'$	n-1

Det er selvfølgelig muligt, ligesom det er gjort i kapitel 5, at bestemme forventede værdier af $\underline{\underline{B}}$ og $\underline{\underline{T}}$ matricerne, også uden, at H_0 er gyldig. Dette skal vi dog ikke komme ind på her.

6.3.2 Tosidet flerdimensional variansanalyse

Vi vil alene betragte en tosidet variansanalyse med 1 observation pr. celle. Vi forudsætter altså, at der foreligger observationer

$$\begin{matrix} \underline{Y}_{11} & , \cdots , & \underline{Y}_{1m} \\ \vdots & & \vdots \\ \underline{Y}_{k1} & , \cdots , & \underline{Y}_{km} \end{matrix} \quad ,$$

der antages p-dimensionalt normalt fordelte med samme dispersionsmatrix $\underline{\underline{\Sigma}}$ og med middelværdier

$$E(\underline{Y}_{ij}) = \underline{\mu}_{ij} = \underline{\mu} + \underline{\alpha}_i + \underline{\beta}_j \ ,$$

hvor parametrene $\underline{\alpha}_i$ og $\underline{\beta}_j$ tilfredsstiller

$$\sum_i \underline{\alpha}_i = \sum_j \underline{\beta}_j = \underline{0} \ .$$

Vi ønsker at teste hypoteserne

$$H_0 : \ \underline{\alpha}_1 = \cdots = \underline{\alpha}_k = \underline{0}$$

mod

$$H_1 : \ \exists i(\underline{\alpha}_i \neq \underline{0})$$

og

$$K_0 : \ \underline{\beta}_1 = \cdots = \underline{\beta}_m = \underline{0}$$

mod

$$K_1 : \ \exists j(\underline{\beta}_j \neq \underline{0}) \ .$$

Vi definerer helt analogt til den endimensionale variansanalyse matricerne

$$\underline{\underline{T}} = \sum_{i=1}^{k} \sum_{j=1}^{m} (\underline{Y}_{ij} - \underline{\overline{Y}}_{..})(\underline{Y}_{ij} - \underline{\overline{Y}}_{..})'$$

$$\underline{\underline{Q}}_1 = \sum_{i=1}^{k} \sum_{j=1}^{m} (\underline{Y}_{ij} - \underline{\overline{Y}}_{i.} - \underline{\overline{Y}}_{.j} + \underline{\overline{Y}}_{..})(\underline{Y}_{ij} - \underline{\overline{Y}}_{i.} - \underline{\overline{Y}}_{.j} + \underline{\overline{Y}}_{..})'$$

$$\underline{\underline{Q}}_2 = m \sum_{i=1}^{k} (\underline{\overline{Y}}_{i.} - \underline{\overline{Y}}_{..})(\underline{\overline{Y}}_{i.} - \underline{\overline{Y}}_{..})'$$

$$\underline{\underline{Q}}_3 = k \sum_{j=1}^{m} (\underline{\overline{Y}}_{.j} - \underline{\overline{Y}}_{..})(\underline{\overline{Y}}_{.j} - \underline{\overline{Y}}_{..})' \ .$$

Her er anvendt den sædvanlige notation

$$\underline{\bar{Y}}_{\cdot\cdot} = \frac{1}{km} \sum_{i=1}^{k} \sum_{j=1}^{m} \underline{Y}_{ij}$$

$$\underline{\bar{Y}}_{i\cdot} = \frac{1}{m} \sum_{j=1}^{m} \underline{Y}_{ij} \quad , \quad i = 1,\cdots,k$$

$$\underline{\bar{Y}}_{\cdot j} = \frac{1}{k} \sum_{i=1}^{k} \underline{Y}_{ij} \quad , \quad j = 1,\cdots,m \ .$$

Man ser, at vi også her har den sædvanlige spaltning af den totale variation

$$\underline{\underline{T}} = \underline{\underline{Q}}_1 + \underline{\underline{Q}}_2 + \underline{\underline{Q}}_3 \ ,$$

d.v.s. den totale variation ($\underline{\underline{T}}$) er spaltet i variationen mellem rækker ($\underline{\underline{Q}}_2$), variationen mellem søjler ($\underline{\underline{Q}}_3$) og restvariationen (vekselvirkningsvariationen) ($\underline{\underline{Q}}_1$).

Der gælder nu

Sætning 6.13 Kvotienttestet på niveau α for test af H_0 med H_1 er givet ved det kritiske område

$$\{\underline{Y}_{11},\cdots,\underline{Y}_{km} \mid \frac{\det(\underline{\underline{q}}_1)}{\det(\underline{\underline{q}}_1 + \underline{\underline{q}}_2)} \leq U(p,k-1,(k-1)(m-1))_\alpha\} \ .$$

Kvotienttestet på niveau α for test af K_0 mod K_1 er givet ved det kritiske område

$$\{\underline{Y}_{11},\cdots,\underline{Y}_{km} \mid \frac{\det(\underline{\underline{q}}_1)}{\det(\underline{\underline{q}}_1 + \underline{\underline{q}}_3)} \leq U(p,m-1,(k-1)(m-1))_\alpha\} \ .$$

Bevis Forbigås. Følger umiddelbart af sætning 6.10. Se fx. Anderson (1958).

Vi samler resultaterne i et sædvanligt variansanalyseskema

Variations-kilde	SAK-matrix	Frihedsgrader	Test-størrelse
Forskelle mellem søjler	$\underline{\underline{Q}}_3 = k \sum_j (\underline{\bar{Y}}_{\cdot j} - \underline{\bar{Y}}_{\cdot\cdot})(\underline{\bar{Y}}_{\cdot j} - \underline{\bar{Y}}_{\cdot\cdot})'$	$m-1$	$\frac{\det(\underline{\underline{Q}}_3)}{\det(\underline{\underline{Q}}_1 + \underline{\underline{Q}}_3)}$
Forskelle mellem rækker	$\underline{\underline{Q}}_2 = m \sum_i (\underline{\bar{Y}}_{i\cdot} - \underline{\bar{Y}}_{\cdot\cdot})(\underline{\bar{Y}}_{i\cdot} - \underline{\bar{Y}}_{\cdot\cdot})'$	$k-1$	$\frac{\det(\underline{\underline{Q}}_2)}{\det(\underline{\underline{Q}}_1 + \underline{\underline{Q}}_2)}$
Residual	$\underline{\underline{Q}}_1 = \sum_i \sum_j (\underline{Y}_{ij} - \underline{\bar{Y}}_{i\cdot} - \underline{\bar{Y}}_{\cdot j} + \underline{\bar{Y}}_{\cdot\cdot})\times (\underline{Y}_{ij} - \underline{\bar{Y}}_{i\cdot} - \underline{\bar{Y}}_{\cdot j} + \underline{\bar{Y}}_{\cdot\cdot})'$	$(k-1)(m-1)$	
Total	$\underline{\underline{T}} = \sum_i \sum_j (\underline{Y}_{ij} - \underline{\bar{Y}}_{\cdot\cdot})(\underline{Y}_{ij} - \underline{\bar{Y}}_{\cdot\cdot})'$	$km-1$	

Matricen $\frac{1}{(k-1)(m-1)}\underline{\underline{Q}}_1$ kan anvendes som et centralt skøn over $\underline{\underline{\Sigma}}$.

Vi giver nu et illustrativt eksempel

Eksempel 6.5 På Landbohøjskolens forsøgsstation Højbakkegård gennemførtes i perioden 1956-1958 som led i et internationalt forsøgsarbejde forsøg med udbytte ved dyrkning af planter. Der blev udført forsøg med 10 typer af planter. De former for udbytte, der havde interesse, var mængderne af

tørstof (dry matter)
grønt (green matter)
kvælstof (nitrogen) .

Hver plantetype blev dyrket i 6 blokke (i.e. jordstykker af forskellig kvalitet). For at reducere datamængden vil vi her indskrænke os til tre planter og til året 1957. Resultaterne af det betragtede forsøg er givet nedenfor

Plante-type	Udbytte-type	Blok nr. 1	2	3	4	5	6
Marchi-giana	Tørstof	9.170	10.683	10.063	8.104	10.018	9.570
	Kvælstof	0.286	0.335	0.315	0.259	0.319	0.304
	Grønt	40.959	47.677	44.950	36.919	45.859	43.838
Kayseri	Tørstof	9.403	10.914	11.018	11.385	13.387	12.848
	Kvælstof	0.285	0.330	0.333	0.339	0.400	0.383
	Grønt	42.475	49.546	50.152	51.718	60.758	58.334
Atlan-tic	Tørstof	11.349	10.971	9.794	8.944	11.715	11.903
	Kvælstof	0.369	0.357	0.319	0.291	0.379	0.386
	Grønt	52.475	50.757	45.151	42.221	55.505	56.364

Udbytte i 1000 kg/ha.

Vi ønsker at analysere, hvordan udbyttet varierer med blokkene, plantetypen og udbyttetypen.

Vi vil først analysere hver udbyttetype for sig og basere analysen på en tosidet variansanalyse. Modellen er

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij} \quad (i = 1,2,3,\ j = 1,\cdots,6) \ ,$$

og vi antager således, at hver observation y_{ij} kan skrives som sum af μ (et niveau), α_i (planteeffekt), β_j (blokeffekt) og ε_{ij} (rest, der er en lille og tilfældigt varierende størrelse).

Betragter vi først tørstof, fås

$$y_{11} = 9.170\ , \quad y_{12} = 10.683,\cdots\ , \quad y_{36} = 11.093\ .$$

Variansanalyseskemaet blev (fundet ved hjælp af SSP-ANOVA)

Source of variation	Sums of squares	Degrees of freedom	Mean squares	F-værdier
A	11.218244	5	2.243648	2.25
B	10.945597	2	5.472798	5.49
AB	9.970109	10	0.997010	
Total	32.133936	17		

Teststørrelsen for hypotesen $\beta_1 = \cdots = \beta_6 = 0$ er

$$F = \frac{s_3^2}{s_1^2} = 2.25 < 3.33 = F_{95\%}(5,10)$$

ɔ: vi kan ikke afvise, at β-erne er lig nul.

Tilsvarende er teststørrelsen for hypotesen $\alpha_1 = \alpha_2 = \alpha_3 = 0$ lig

$$F = \frac{s_2^2}{s_1^2} = 5.49 > 4.10 = F_{95\%}(2,10) \quad .$$

Vi kan således afvise på 5% niveauet, at α-erne er alle lig nul. Men vi bemærker, at

$$F_{97.5\%}(2,10) = 5.46 \quad ,$$

således at der ikke er signifikans på ca. 2.5% niveauet.

Udføres de tilsvarende beregninger på kvælstofudbyttet, fås, idet vi som observationer anvender $y'_{ij} = y_{ij} \cdot 1000$:

Source of variation	Sums of squares	Degrees of freedom	Mean squares	F-værdier
A	10802.27734	5	2160.45532	2.60
B	8030.77734	2	4015.38867	4.83
AB	8310.55469	10	831.05542	
Total	27143.60938	17		

Her får vi ligeledes, at der ikke er forskel på blokkene, men der er muligvis forskel på planterne. Denne forskel er dog ikke signifikant på niveau 2.5%.

De tilsvarende beregninger på grøntudbyttet blev, idet vi fremdeles anvender kodede observationer ($y'_{ij} = 1000\ y_{ij}$)

Source of variation	Sums of squares	Degrees of freedom	Mean squares	F-værdier
A	261702416	5	52340480	2.75
B	260173824	2	130086912	6.83
AB	190600448	10	19060032	
Total	712476672	17		

Her får vi igen, at der ikke er forskel på blokkene. Vi får også en forskel på planterne på 5% niveauet, men ikke på 1% niveauet, idet

$$F_{99\%}(2,10) = 7.56 \quad .$$

Vi ser således, at de tre former for udbytte viser nogenlunde samme form for variation: der er ikke forskel på blokkene, men der er forskel på planterne, der dog ikke er signifikante på et lille niveau.

Nu er de tre former for udbytter stærkt afhængige. Det var derfor at forvente, at variansanalyserne ville give nogenlunde ensartede resultater, og det vil følgelig være interessant at undersøge variationen i udbyttet, når vi tager hensyn til denne afhængighed. En sådan analyse kan gennemføres ved en 3-dimensional tosidet variansanalyse, d.v.s. vi arbejder med modellen

$$\underline{Y}_{ij} = \underline{\mu} + \underline{\alpha}_i + \underline{\beta}_j + \underline{\varepsilon}_{ij} \quad , \quad i = 1,2,3 \, , \quad j = 1,\cdots,6 \quad ,$$

hvor

$$\underline{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} \quad , \quad \underline{\alpha}_i = \begin{pmatrix} \alpha_{1i} \\ \alpha_{2i} \\ \alpha_{3i} \end{pmatrix} \quad , \quad \underline{\beta}_j = \begin{pmatrix} \beta_{1j} \\ \beta_{2j} \\ \beta_{3j} \end{pmatrix} \quad ,$$

og observationerne er

$$\underline{Y}_{ij} = \begin{pmatrix} \text{grøntindhold} & \text{i plante i på blok j} \\ \text{kvælstofindhold} & \text{-} \quad \text{"} \quad \text{-} \\ \text{tørstofindhold} & \text{-} \quad \text{"} \quad \text{-} \end{pmatrix} \, .$$

De realiserede udfald er

$$\underline{Y}_{11} = \begin{pmatrix} 0.959 \\ 0.286 \\ 9.170 \end{pmatrix} \quad , \cdots, \quad \underline{Y}_{36} = \begin{pmatrix} 56.364 \\ 0.386 \\ 11.903 \end{pmatrix} \, .$$

Vi slår således de tre variansanalysemodeller ovenfor sammen i én.

Med notationen fra p. 9.31 findes de realiserede udfald af matricerne $\underline{\underline{Q}}_1$, $\underline{\underline{Q}}_2$ og $\underline{\underline{Q}}_3$ til

$$\underline{\underline{q}}_2 = \begin{bmatrix} 260.18359 & & \\ 1.38547 & 0.00803 & \\ 52.37032 & 0.26262 & 10.94564 \end{bmatrix}$$

$$\underline{\underline{q}}_3 = \begin{bmatrix} 261.70239 & & \\ 1.67129 & 0.01080 & \\ 53.97473 & 0.34801 & 11.21827 \end{bmatrix}$$

$$\underline{\underline{q}}_1 = \begin{bmatrix} 190.59937 & & \\ 1.25512 & 0.00831 & \\ 43.45444 & 0.28667 & 9.97013 \end{bmatrix}$$

Matricerne er fundet ved hjælp af BMD-programmet BMDX69. Stadig ved hjælp af dette program findes

Scource	Log(Generalized variance)	U-statistic	Degrees of freedom			Approximate F-statistic	Degrees of freedom	
I	-1.89908	0.003315	3	2	10	43.6455	6	16.00
J	-4.84194	0.062894	3	5	10	2.5843	15	22.49
Full model	-7.60824							

Her svarer I til variation mellem planter og J til variation mellem blokke.

Den (her eksakte) F-teststørrelse for test af hypotesen $\underline{\alpha}_1 = \underline{\alpha}_2 = \underline{\alpha}_3 = 0$, d.v.s. hypotesen, at alle planter er ens, er 43.6. Antallet af frihedsgrader er (6,16). Da

$$F(6,16)_{0.9995} = 7.74 \ ,$$

er der altså tale om en meget kraftig forkastelse af hypotesen.

Da

$$F(15,22)_{0.975} = 2.50 \ ,$$

ser vi, at hypotesen om ens blokke lige netop bliver forkastet på niveau $\alpha = 2.5\%$.

Konklusionen på den flerdimensionale variansanalyse er derfor, at der er en meget tydelig forskel på udbytterne for de tre plantetyper. Det synes derimod mere tvivlsomt, om der er forskelle på blokkene.

Vi bemærker her en forskel fra de tre endimensionale analyser. Der var der kun tale om moderate eller ingen signifikans for hypotesen om ens planteudbytter. Man får således forskellige resultater frem ved at betragte den simultane analyse i stedet for de tre marginale.

□

6.4 Tests vedrørende dispersionsmatricer

I dette afsnit skal vi kort omtale nogle tests for hypoteser om dispersionsmatricer, dels svarende til en hypotese om, at en dispersionsmatrix har en given struktur eller er lig en given matrix, og dels svarende til en hypotese om, at flere dispersionsmatricer er ens.

6.4.1 Tests vedrørende en enkelt dispersionsmatrix

Vi skal her først angive et test for, at k grupper af normalt fordelte variable er uafhængige. Vi betragter altså et $\underline{X} \in N_p(\underline{\mu},\underline{\Sigma})$, og vi inddeler $\underline{X}$ i k komponenter med dimensionerne $p_1,\cdots,p_k$, d.v.s.:

$$\underline{X} = \begin{bmatrix} \underline{X}_1 \\ \vdots \\ \underline{X}_k \end{bmatrix} .$$

Den tilsvarende spaltning af parametrene er

$$\underline{\mu} = \begin{bmatrix} \underline{\mu}_1 \\ \vdots \\ \underline{\mu}_k \end{bmatrix}$$

og

$$\underline{\underline{\Sigma}} = \begin{bmatrix} \underline{\underline{\Sigma}}_{11} & \cdots & \underline{\underline{\Sigma}}_{1k} \\ \vdots & & \vdots \\ \underline{\underline{\Sigma}}_{k1} & \cdots & \underline{\underline{\Sigma}}_{kk} \end{bmatrix} .$$

Vores hypotese er nu, at $\underline{X}_1, \cdots, \underline{X}_k$ er uafhængige, d.v.s. at dispersionsmatricen har formen

$$\underline{\underline{\Sigma}} = \underline{\underline{\Sigma}}_0 = \begin{bmatrix} \underline{\underline{\Sigma}}_{11} & \cdots & \underline{\underline{0}} \\ \vdots & & \vdots \\ \underline{\underline{0}} & \cdots & \underline{\underline{\Sigma}}_{kk} \end{bmatrix} .$$

Definerer vi $\hat{\underline{\underline{\Sigma}}}$ beregnet på basis af n realisationer af $\underline{X}$ på sædvanlig måde, og spaltes $\hat{\underline{\underline{\Sigma}}}$ analogt til spaltningen af $\underline{\underline{\Sigma}}$, har vi

Sætning 6.16 Vi betragter ovenstående situation og sætter

$$V = \frac{\det(\hat{\underline{\underline{\Sigma}}})}{\prod_{i=1}^{k} \det(\hat{\underline{\underline{\Sigma}}}_{ii})} .$$

Da er kvotienttestet for test af hypotesen $\underline{\underline{\Sigma}} = \underline{\underline{\Sigma}}_0$ givet ved det kritiske område

$$\{V \leq v_\alpha\} \quad .$$

Ved fastlæggelse af grænsen i det kritiske område kan benyttes, at

$$P\{-m \log V \leq v\} \simeq P\{\chi^2(f) \leq v\} + \frac{\gamma_2}{m^2}\left[P\{\chi^2(f+4) \leq v\} - P\{\chi^2(f) \leq v\}\right] ,$$

hvor

$$m = n - \frac{3}{2} - \frac{p^3 - \Sigma p_i^3}{3(p^2 - \Sigma p_i^2)}$$

$$\gamma_2 = \frac{p^4 - \Sigma p_i^4}{48} - \frac{5(p^2 - \Sigma p_i^2)}{96} - \frac{(p^3 - \Sigma p_i^3)^2}{72(p^2 - \Sigma p_i^2)} \quad .$$

Hvis $k = 2$, er V fordelt som $U(p_1, p_2, n-1-p_2)$.

Bevis Forbigås. Se fx. Anderson (1958).

I ovenstående situation så vi på et test for, at en dispersionsmatrix havde en bestemt struktur. Vi skal nu vende os mod et test for en hypotese, at en dispersionsmatrix er proportional med en given matrix. Vi formulerer kort resultatet i

Sætning 6.17 Vi betragter uafhængige observationer $\underline{X}_1, \cdots, \underline{X}_n$ med $\underline{X}_i \in N_p(\underline{\mu}, \underline{\underline{\Sigma}})$, og vi sætter

$$\underline{\underline{A}} = \Sigma(\underline{X}_i - \overline{\underline{X}})(\underline{X}_i - \overline{\underline{X}})' \quad .$$

Kvotienttesstørrelsen for test af $H_0 : \underline{\underline{\Sigma}} = \sigma^2 \underline{\underline{\Sigma}}_0$, hvor $\underline{\underline{\Sigma}}_0$ er kendt og σ^2 ukendt, mod alle alternativer er

$$W = \frac{\left[\det(\underline{\underline{A}}\ \underline{\underline{\Sigma}}_0^{-1})\right]^{n/2}}{\left[\mathrm{tr}\ \underline{\underline{A}}\ \underline{\underline{\Sigma}}_0^{-1}/p\right]^{pn/2}} \quad .$$

Ved fastlæggelse af det kritiske område kan benyttes, at

$$P\{-(n-1)\rho \log W \leq z\}$$

$$\simeq P\{\chi^2(f) \leq z\} + \omega_2 \left[P\{\chi^2(f+4) \leq z\} - P\{\chi^2(f) \leq z\}\right] ,$$

hvor

$$\rho = 1 - \frac{2p^2 + p + 2}{6p(n-1)}$$

$$f = \frac{1}{2}p(p+1) - 1$$

$$\omega_2 = \frac{(p+2)(p-1)(p-2)(2p^3 + 6p^2 + 3p + 2)}{288\, p^2\, n^2\, \rho^2} .$$

Bevis Forbigås. Se fx. Anderson (1958).

Endelig skal vi betragte situationen, hvor vi ønsker at teste, at en dispersionsmatrix er lig en given matrix. Der gælder

Sætning 6.18 Vi betragter indbyrdes uafhængige observationer $\underline{X}_1, \cdots, \underline{X}_n$ med $\underline{X}_i \in N_p(\underline{\mu}, \underline{\underline{\Sigma}})$, og vi sætter

$$\underline{\underline{A}} = \sum_{i=1}^{n} (\underline{X}_i - \overline{\underline{X}})(\underline{X}_i - \overline{\underline{X}})' .$$

Kvotienteststørrelsen for test af $H_0 : \underline{\underline{\Sigma}} = \underline{\underline{\Sigma}}_0$, hvor $\underline{\underline{\Sigma}}_0$ er kendt, mod alle alternativer, er

$$\lambda_1 = \left(\frac{e}{n}\right)^{pn/2} \left[\det(\underline{\underline{A}}\, \underline{\underline{\Sigma}}_0^{-1})\right]^{n/2} \exp\left(-\frac{1}{2}\mathrm{tr}(\underline{\underline{A}}\, \underline{\underline{\Sigma}}_0^{-1})\right) .$$

Ved fastlæggelse af det kritiske område kan benyttes, at

$$P\{-2 \log \lambda_1 \leq v\} \simeq P\{\chi^2(\tfrac{1}{2}p(p+1)) \leq v\} .$$

Bevis Forbigås. Se fx. Anderson (1958).

6.4.2 Test for, at flere dispersionsmatricer er ens

Vi betragter i dette afsnit problemet med at teste forudsætningen om ens dispersionsmatricer i Hotellings tostikprøve-situation og i de flerdimensionale variansanalyser.

Vi antager altså, at der foreligger uafhængige observationer

$$\begin{array}{cccc} \underline{X}_{11}, \cdots, \underline{X}_{1n_1}, & \underline{X}_{1j} \in N_p(\underline{\mu}_1, \underline{\underline{\Sigma}}_1) \\ \vdots \qquad\qquad \vdots & \\ \underline{X}_{k1}, \cdots, \underline{X}_{kn_k}, & \underline{X}_{kj} \in N_p(\underline{\mu}_k, \underline{\underline{\Sigma}}_k) \end{array},$$

og vi ønsker at teste hypotesen

$$H_0 : \underline{\underline{\Sigma}}_1 = \cdots = \underline{\underline{\Sigma}}_1 \quad \text{mod} \quad H_1 : \exists i,j : \underline{\underline{\Sigma}}_i \neq \underline{\underline{\Sigma}}_j .$$

Vi sætter

$$n = \Sigma n_i ,$$

$$\underline{\underline{A}}_i = \sum_{j=1}^{n_i} (\underline{X}_{ij} - \overline{\underline{X}}_i)(\underline{X}_{ij} - \overline{\underline{X}}_i)' ,$$

og

$$\underline{\underline{A}} = \sum_{i=1}^{k} \underline{\underline{A}}_i ,$$

jvf. afsnit 6.3.1, hvor betegnelsen $\underline{\underline{W}}$ er anvendt i stedet for $\underline{\underline{A}}$.

Vi har da

Sætning 6.19 Som teststørrelse for test af H_0 mod H_1 kan benyttes

$$W_1 = \frac{\prod_{i=1}^{k} \left[\det(\underline{\underline{A}}_i)\right]^{(n_i-1)/2}}{\left[\det \underline{\underline{A}}\right]^{(n-k)/2}} \cdot \frac{(n-k)^{p(n-k)/2}}{\prod_{i=1}^{k} (n_i-1)^{p(n_i-1)/2}} .$$

Det kritiske område er af formen

$$\{W_1 \leq w_\alpha\} ,$$

og ved fastlæggelse af dette kan benyttes, at

$$P\{-2\rho \log W_1 \leq z\}$$

$$\approx P\{\chi^2(f) \leq z\} + \omega_2 \left[P\{\chi^2(f+4) \leq z\} - P\{\chi^2(f) \leq z\}\right] ,$$

hvor

$$f = \tfrac{1}{2}(k-1)p(p+1) ,$$

$$\rho = 1 - \left(\sum_i \frac{1}{n_i} - \frac{1}{n}\right) \frac{2p^2+3p-1}{6(p+1)(q-1)} ,$$

$$\omega_2 = \frac{1}{48\rho^2}\, p(p+1)\left[(p-1)(p+2)\left(\sum_i \frac{1}{n_i^2} - \frac{1}{n^2}\right) - 6(k-1)(1-\rho)^2\right] .$$

Bevis Forbigås. Se fx. Anderson (1958).

Reference til kapitel 6

Anderson, T.W.: An Introduction to Multivariate Statistical Analysis. John Wiley & Sons, New York 1958.

Cooley, W. W. & P. R. Lohnes: Multivariate Data Analysis. John Wiley & Sons, New York 1971.

KAPITEL 7

Diskriminantanalyse

I dette afsnit skal vi beskæftige os med det problem at klassificere et individ i en af 2 (eller flere) kendte populationer på basis af målinger af nogle karakteristika ved individet.

Vi betragter nu først problemet med at skelne ("diskriminere") mellem 2 grupper (klasser).

7.1 Diskrimination mellem 2 populationer

7.1.1 Bayes- og minimaxløsninger

Vi betragter populationerne π_1 og π_2 og ønsker at afgøre, om et forelagt individ hører hjemme i gruppe 1 eller gruppe 2. Vi foretager målinger af p forskellige egenskaber ved individet og får derved resultatet

$$\underline{X} = \begin{pmatrix} X_1 \\ \cdot \\ \cdot \\ \cdot \\ X_p \end{pmatrix} .$$

Hvis individet stammer fra π_1, er frekvensfunktionen for $\underline{X}$ $f_1(\underline{x})$, og hvis det stammer fra π_2, er den $f_2(\underline{x})$.

Lad os endvidere antage, at der er givet en tabsfunktion L:

		Vælger	
		π_1	π_2
Tilstand:	π_1	0	L(1,2)
	π_2	L(2,1)	0

Vi regner ikke med, at der er et tab, hvis vi tager den rigtige beslutning.

I visse situationer ved man også nogenlunde, hvad a priori sandsynligheden er for at få et individ fra hver af grupperne, d.v.s. der er givet en a priorifordeling g:

$$g(\pi_1) = p_1, \quad g(\pi_2) = p_2 .$$

Vi søger nu en beslutningsfunktion d: $R^p \to \{\pi_1, \pi_2\}$. d defineres ved

$$d(\underline{x}) = d_{R_1}(\underline{x}) = \begin{cases} \pi_1 \text{ hvis } \underline{x} \in R_1 \\ \pi_2 \text{ hvis } \underline{x} \in R_2 = \complement R_1 \end{cases} .$$

Vi inddeler altså R^p i 2 områder R_1 og R_2. Hvis vort udfald ligger i R_1, vælger vi π_1, og hvis det ligger i R_2, vælger vi π_2 .

Hvis vi har en a priorifordeling, definerer vi a posteriorifordelingen k ved

$$k(\pi_i|\underline{x}) = \frac{f_i(\underline{x})g(\pi_i)}{p_1 f_1(\underline{x}) + p_2 f_2(\underline{x})} = \frac{p_i f_i(\underline{x})}{p_1 f_1(\underline{x}) + p_2 f_2(\underline{x})} ,$$

jvf. p. 6.6 i bind 1.

Det forventede tab i denne fordeling er

$$E_{\underline{x}}(L(\pi,d_{R_1}(\underline{x}))) = L(\pi_1,d_{R_1}(\underline{x}))k(\pi_1|\underline{x}) + L(\pi_2,d_{R_1}(\underline{x}))k(\pi_2|\underline{x})$$

$$= \begin{cases} L(\pi_2,\pi_1)k(\pi_2|\underline{x}) & , \quad \underline{x} \in R_1 \\ L(\pi_1,\pi_2)k(\pi_1|\underline{x}) & , \quad \underline{x} \in R_2 \end{cases} .$$

Bayesløsningen defineres ved, at vi skal minimalisere denne størrelse for ethvert $\underline{x}$ (p. 6.9 i bind 1), d.v.s. vi må definere R_1 ved

$$\underline{x} \in R_1 \Leftrightarrow L(2,1)k(\pi_2|\underline{x}) \leq L(1,2)k(\pi_1|\underline{x})$$

$$\Leftrightarrow \frac{L(1,2)\ f_1(\underline{x})\ p_1}{L(2,1)\ f_2(\underline{x})\ p_2} \geq 1$$

$$\Leftrightarrow \frac{f_1(\underline{x})}{f_2(\underline{x})} \geq \frac{L(2,1)}{L(1,2)}\ \frac{p_2}{p_1} .$$

Vi samler disse overvejelser i

Sætning 7.1 Bayesløsningen til klassifikationsproblemet er givet ved området

$$R_1 = \{ \underline{x} \mid \frac{f_1(\underline{x})}{f_2(\underline{x})} \geq \frac{L(2,1)}{L(1,2)}\ \frac{p_2}{p_1} \} .$$

Bemærkning Dette resultat er præcis det samme, som står anført i sætning 5, kap. 6 i bind 1.

Hvis vi ikke har en a priori fordeling, kan vi bestemme en minimax strategi, i.e. bestemme et R_1, således at den maksimale risiko minimaliseres. Risikoen er (jvf. p. 6.3, bind 1)

$$R(\pi_1,d_{R_1}) = E_{\pi_1}L(\pi_1,d_{R_1}(\underline{X})) = L(1,2)P\{\underline{X}\in R_2|\pi_1\} .$$

$$R(\pi_2,d_{R_1}) = E_{\pi_2}L(\pi_2,d_{R_1}(\underline{X})) = L(2,1)P\{\underline{X}\in R_1|\pi_2\} .$$

Man kan nu vise (se e.g. beviset for sætning 4, kap. 6 i bind 1)

Sætning 7.2 Minimaxløsningen til klassifikationsproblemet er givet ved området

$$R_1 = \{ \underline{x} \mid \frac{f_1(\underline{x})}{f_2(\underline{x})} \geq c \} ,$$

hvor c bestemmes ved

$$L(1,2)P\{ \frac{f_1(\underline{X})}{f_2(\underline{X})} < c | \pi_1 \} = L(2,1)P\{ \frac{f_1(\underline{X})}{f_2(\underline{X})} \geq c | \pi_2 \} .$$

Bemærkning Relationen til bestemmelse af c kan skrives

$$L(1,2) \cdot (\text{sandsynligheden for misklassifikation hvis } \pi_1 \text{ er sand})$$

$$= L(2,1) \cdot (\text{sandsynligheden for misklassifikation hvis } \pi_2 \text{ er sand}) .$$

Da den ene åbenbart er voksende og den anden aftagende i c, er det klart, at vi netop får minimaliseret den maksimale risiko, når der er lighedstegn. Hvis vi ikke har nogen ideer om tabenes størrelse, kan vi sætte dem begge til 1. Minimaxløsningen giver os da det område, der minimaliserer den maksimale sandsynlighed for misklassifikation.

Vi betragter nu det vigtige specialtilfælde, hvor f_1 og f_2 er normalfordelinger.

7.1.2 Diskrimination mellem 2 normale populationer

Hvis f_1 og f_2 er normale med samme dispersionsmatrix, har vi

Sætning 7.3 Lad $\pi_1 \simeq N(\underline{\mu}_1, \underline{\underline{\Sigma}})$ og $\pi_2 \simeq N(\underline{\mu}_2, \underline{\underline{\Sigma}})$. Da gælder

$$\frac{f_1(\underline{x})}{f_2(\underline{x})} \geq c \Leftrightarrow \underline{x}'\underline{\underline{\Sigma}}^{-1}(\underline{\mu}_1-\underline{\mu}_2) - \tfrac{1}{2}\underline{\mu}_1'\underline{\underline{\Sigma}}^{-1}\underline{\mu}_1 + \tfrac{1}{2}\underline{\mu}_2'\underline{\underline{\Sigma}}^{-1}\underline{\mu}_2 \geq \log c \ .$$

Bevis Vi indfører det indre produkt $(\cdot|\cdot)$ og normen $||\ \ ||$ ved

$$(\underline{x}|\underline{y}) = \underline{x}'\underline{\underline{\Sigma}}^{-1}\underline{y}$$

og

$$||x||^2 = (\underline{x}|\underline{x}) \ .$$

Vi har da

$$f_i(\underline{x}) = \frac{1}{\sqrt{2\pi}^p \sqrt{\det \underline{\underline{\Sigma}}}} \exp(-\tfrac{1}{2}||\underline{x} - \underline{\mu}_i||^2) \ .$$

Heraf fås umiddelbart

$$\frac{f_1(\underline{x})}{f_2(\underline{x})} \geq c \Leftrightarrow \log \frac{f_1(\underline{x})}{f_2(\underline{x})} \geq \log c$$

$$\Leftrightarrow - ||\underline{x} - \underline{\mu}_1||^2 + ||\underline{x} - \underline{\mu}_2||^2 \geq 2 \log c$$

$$\Leftrightarrow - (\underline{x}-\underline{\mu}_1|\underline{x}-\underline{\mu}_1)+(\underline{x}-\underline{\mu}_2|\underline{x}-\underline{\mu}_1) \geq 2 \log c$$

$$\Leftrightarrow 2\ (\underline{x}|\underline{\mu}_1)-2(\underline{x}|\underline{\mu}_2)-(\underline{\mu}_1|\underline{\mu}_1)+(\underline{\mu}_2|\underline{\mu}_2) \geq 2 \log c$$

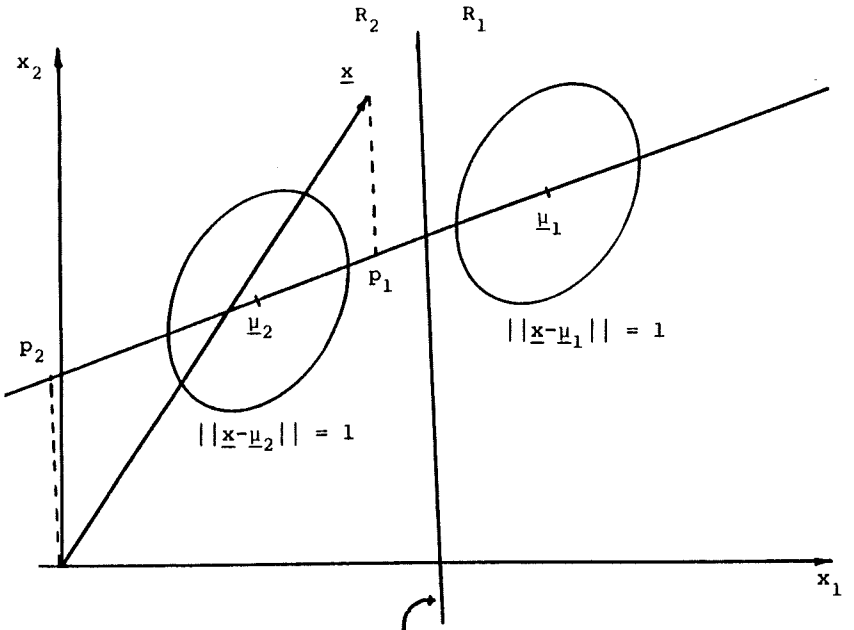
$$\Leftrightarrow 2\ (\underline{x}|\underline{\mu}_1-\underline{\mu}_2)-(\underline{\mu}_1|\underline{\mu}_1)+(\underline{\mu}_2|\underline{\mu}_2) \geq 2 \log c \ .$$

Ved at benytte sammenhængen mellem $(|)$ og $\underline{\underline{\Sigma}}^{-1}$ fås sætningen umiddelbart.

□

Bemærkning Udtrykket $\dfrac{f_1(\underline{x})}{f_2(\underline{x})} \geq c$ ses at definere en delmængde af

R^p, der er afgrænset af en hyperplan (for p = 2 en ret linie og for p = 3 en plan)



$$\underline{x}'\underline{\underline{\Sigma}}^{-1}(\underline{\mu}_1-\underline{\mu}_2) - \tfrac{1}{2}\underline{\mu}_1'\underline{\underline{\Sigma}}^{-1}\underline{\mu}_1 + \tfrac{1}{2}\underline{\mu}_2'\underline{\underline{\Sigma}}^{-1}\underline{\mu}_2 - \log c = 0 \ .$$

Vektoren $\vec{p_1p_2}$ betegner den ortogonale projektion (NB! ortogonal med hensyn til $\underline{\underline{\Sigma}}^{-1}$) af $\underline{x}$ på linien, der forbinder $\underline{\mu}_1$ og $\underline{\mu}_2$. (Det kan vises, at hældningen af projektionslinierne m.v. er lig hældningen af ellipse (ellipsoide) tangenterne i de punkter, hvor de skærer linien ($\underline{\mu}_1$, $\underline{\mu}_2$). Da længden af en projektion af en vektor er lig det indre produkt mellem vektoren og en enhedsvektor på linien, ser vi, at vi klassificerer observationen som stammende fra π_1, netop hvis projektionen af $\underline{x}$ er tilstrækkelig lang (regnet med fortegn). Ellers klassificerer vi observationen som stammende fra π_2.

Funktionen

$$\underline{x}'\underline{\underline{\Sigma}}^{-1}(\underline{\mu}_1-\underline{\mu}_2) - \tfrac{1}{2}\underline{\mu}_1'\underline{\underline{\Sigma}}^{-1}\underline{\mu}_1 + \tfrac{1}{2}\underline{\mu}_2'\underline{\underline{\Sigma}}^{-1}\underline{\mu}_2 - \log c$$

kaldes diskriminatoren eller diskriminantfunktionen.

Vi har altså, at diskriminatoren er den lineære afbildning, der - efter addition af passende konstanter - minimaliserer det forventede tab (Bayes-situationen) eller misklassifikationssandsynlighederne (minimax-situationen).

Vi skal nu - for at gøre læseren fortrolig med indholdet i begrebet en diskriminator - give en lidt anden tolkning af denne. Sætter vi

$$\underline{\delta} = \underline{\underline{\Sigma}}^{-1}(\underline{\mu}_1 - \underline{\mu}_2) \ ,$$

har vi følgende

Sætning 7.4 Vektoren $\underline{\delta}$ har den egenskab, at den maksimaliserer funktionen

$$\varphi(\underline{d}) = \frac{[E_1(\underline{X}'\underline{d}) - E_2(\underline{X}'\underline{d})]^2}{V(\underline{X}'\underline{d})} = \frac{[(\underline{\mu}_1 - \underline{\mu}_2)'\underline{d}]^2}{\underline{d}'\underline{\underline{\Sigma}}\,\underline{d}} \ .$$

Bevis Beviset er ikke særlig interessant, men relativt simpelt. Da vi umiddelbart ser, at $\varphi(k\cdot\underline{d}) = k\cdot\varphi(\underline{d})$, kan vi bestemme ekstremaer for φ ved at bestemme ekstremaer for tælleren under betingelsen

$$\underline{d}'\,\underline{\underline{\Sigma}}\,\underline{d} = 1 \ .$$

Vi indfører en Lagrange multiplikator λ og søger maksimum af

$$\psi(\underline{d}) = [(\underline{\mu}_1 - \underline{\mu}_2)'\underline{d}]^2 - \lambda(\underline{d}'\underline{\underline{\Sigma}}\,\underline{d} - 1) \ .$$

Nu er

$$\frac{\partial \psi}{\partial \underline{d}} = 2(\underline{\mu}_1 - \underline{\mu}_2)(\underline{\mu}_1 - \underline{\mu}_2)'\underline{d} - 2\,\lambda\,\underline{\underline{\Sigma}}\,\underline{d} \ .$$

Sættes denne lig 0, fås

$$(\underline{\mu}_1 - \underline{\mu}_2)(\underline{\mu}_1 - \underline{\mu}_2)'\underline{d} = \lambda \, \underline{\underline{\Sigma}} \, \underline{d} \ ,$$

d.v.s.

$$\underline{d} = \frac{(\underline{\mu}_1 - \underline{\mu}_2)'\underline{d}}{\lambda} \, \underline{\underline{\Sigma}}^{-1}(\underline{\mu}_1 - \underline{\mu}_2) = k \cdot \delta \ ,$$

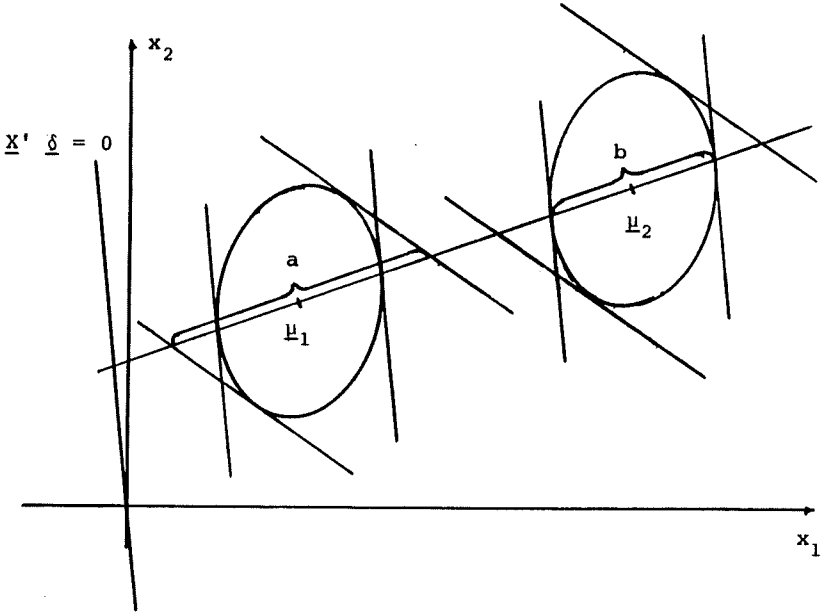
hvor k er en skalar.

□

Bemærkning Sætningens indhold er, at den ved $\underline{\delta}$ bestemte lineære funktion

$$\underline{X}' \, \underline{\delta} = \delta_1 X_1 + \cdots + \delta_p X_p \ ,$$

er den afbildning, der "fjerner" π_1 og π_2 mest fra hinanden, eller - udtrykt i et variansanalysesprog - den afbildning, der maksimaliserer "variansen" mellem populationerne divideret med den totale varians.



Det geometriske indhold af sætningen er søgt anskueliggjort i ovenstående figur, hvor

b: projektionen af ellipsen på linien $\underline{\mu}_1$, $\underline{\mu}_2$ efter retningen bestemt ved $\underline{x}' \underline{\delta} = 0$

a: projektionen af ellipsen på linien $\underline{\mu}_1$, $\underline{\mu}_2$ efter anden retning.

Det fremgår, at den ved $\underline{\delta}$ bestemte projektion på linien, der forbinder $\underline{\mu}_1$ og $\underline{\mu}_2$, er den, der "fjerner" projektionerne af konturellips(oid)erne hørende til de to populationers fordelinger mest muligt fra hinanden.

Vi anfører nu en sætning, som er særdeles nyttig ved bestemmelse af misklassifikationssandsynligheder.

Sætning 7.5 Vi betragter det i sætning 7.3 forekommende kriterium

$$Z = \underline{X}' \underline{\underline{\Sigma}}^{-1} (\underline{\mu}_1 - \underline{\mu}_2) - \tfrac{1}{2} \underline{\mu}_1' \underline{\underline{\Sigma}}^{-1} \underline{\mu}_1 + \tfrac{1}{2} \underline{\mu}_2' \underline{\underline{\Sigma}}^{-1} \underline{\mu}_2 \ .$$

Om dette gælder

$$Z \in \begin{cases} N(+\tfrac{1}{2}||\underline{\mu}_1 - \underline{\mu}_2||^2, \ ||\underline{\mu}_1 - \underline{\mu}_2||^2), & \text{hvis } \pi_1 \text{ sand} \\ N(-\tfrac{1}{2}||\underline{\mu}_1 - \underline{\mu}_2||^2, \ ||\underline{\mu}_1 - \underline{\mu}_2||^2), & \text{hvis } \pi_2 \text{ sand} \end{cases} .$$

Bevis Beviset er helt ligefremt. Lad os e.g. betragte tilfældet π_1 sand. Vi har da $E(\underline{X}) = \underline{\mu}_1$ og dermed

$$\begin{aligned} E(Z) &= \underline{\mu}_1' \underline{\underline{\Sigma}}^{-1} (\underline{\mu}_1 - \underline{\mu}_2) - \tfrac{1}{2} \underline{\mu}_1' \underline{\underline{\Sigma}}^{-1} \underline{\mu}_1 + \tfrac{1}{2} \underline{\mu}_2' \underline{\underline{\Sigma}}^{-1} \underline{\mu}_2 \\ &= \tfrac{1}{2} (\underline{\mu}_1 - \underline{\mu}_2)' \underline{\underline{\Sigma}}^{-1} (\underline{\mu}_1 - \underline{\mu}_2) \\ &= \tfrac{1}{2} ||\underline{\mu}_1 - \underline{\mu}_2||^2 \ . \end{aligned}$$

$$V(Z) = (\underline{\mu}_1-\underline{\mu}_2)'\underline{\underline{\Sigma}}^{-1}\underline{\underline{\Sigma}}\underline{\underline{\Sigma}}^{-1}(\underline{\mu}_1-\underline{\mu}_2)$$

$$= (\underline{\mu}_1-\underline{\mu}_2)'\underline{\underline{\Sigma}}^{-1}(\underline{\mu}_1-\underline{\mu}_2)$$

$$= ||\underline{\mu}_1-\underline{\mu}_2||^2 .$$

Resultatet vedrørende π_2 vises ganske analogt.

□

Vi ser nu på en række eksempler.

Eksempel 7.1 Vi betragter det tilfælde, hvor

$$\pi_1 \leftrightarrow N(\begin{pmatrix}4\\2\end{pmatrix}, \begin{pmatrix}1 & 1\\1 & 2\end{pmatrix})$$

$$\pi_2 \leftrightarrow N(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & 1\\1 & 2\end{pmatrix}) ,$$

og vi vil bestemme en "bedste" diskriminatorfunktion. Da der intet er oplyst om a priori sandsynligheder og lignende, vil vi bestemme den funktion, der svarer til, at konstanten c i sætning 7.3 er 1. Da

$$\begin{pmatrix}1 & 1\\1 & 2\end{pmatrix}^{-1} = \begin{pmatrix}2 & -1\\-1 & 1\end{pmatrix} ,$$

får vi følgende funktion

$$(x_1x_2)\begin{pmatrix}2 & -1\\-1 & 1\end{pmatrix}\begin{pmatrix}3\\1\end{pmatrix} - \tfrac{1}{2}(2\cdot16+1\cdot4-2\cdot8) + \tfrac{1}{2}(2\cdot1+1\cdot1-2\cdot1) = 0$$

eller

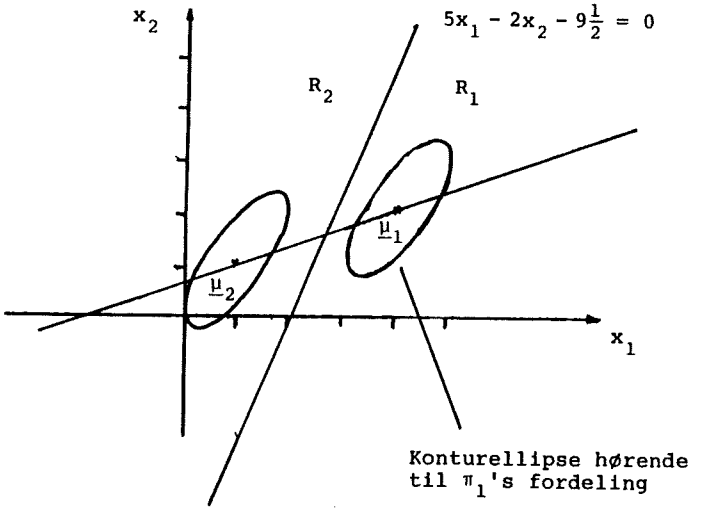
$$5x_1-2x_2-9\tfrac{1}{2} = 0 .$$

Indsætter vi et vilkårligt punkt, e.g. $\begin{pmatrix}5\\6\end{pmatrix}$, fås

$$5\cdot5-2\cdot6-9\tfrac{1}{2} = 3\tfrac{1}{2} > 0.$$

Dette punkt bliver altså klassificeret som stammende fra π_1.

Vi har skitseret situationen i nedenstående figur



□

Hvis vi har en tabsfunktion, bliver fremgangsmåden anderledes, som det fremgår af

Eksempel 7.2 Lad os antage, at vi har visse tab knyttet til de forskellige beslutninger:

		vælge	
		π_1	π_2
natur	π_1	0	2
	π_2	1	0

Da der ingen a priori sandsynligheder foreligger, bestemmer vi minimaxløsningen. Vi får brug for

$$||\underline{\mu}_1 - \underline{\mu}_2||^2 = 2 \cdot 9 + 1 \cdot 1 - 2 \cdot 3 \cdot 1 = 13 \ .$$

Af sætning 7.2 følger, at vi skal bestemme c, så

$$2 \cdot P\left\{\frac{f_1(\underline{X})}{f_2(\underline{X})} < c \mid \pi_1\right\} = P\left\{\frac{f_1(\underline{X})}{f_2(\underline{X})} \geq c \mid \pi_2\right\}$$

$$\Leftrightarrow 2 \cdot P\{Z < \log c \mid \pi_1\} = P\{Z \geq \log c \mid \pi_2\}$$

$$\Leftrightarrow 2 \cdot P\{N(\tfrac{1}{2}13, 13) < \log c\} = P\{N(-\tfrac{1}{2}13, 13) \geq \log c\}$$

$$\Leftrightarrow 2 \cdot P\left\{N(0,1) < \frac{\log c - 6.5}{\sqrt{13}}\right\} = P\left\{N(0,1) \geq \frac{\log c + 6.5}{\sqrt{13}}\right\} .$$

Ved at prøve med forskellige værdier af c indses, at

$$c \simeq 0.5617 .$$

Med denne værdi er misklassifikationssandsynlighederne

$$\text{Hvis } \pi_1 \text{ sand: } P\left\{N(0,1) < \frac{\log 0.5617 - 6.5}{\sqrt{13}}\right\} \simeq 0.025 .$$

$$\text{Hvis } \pi_2 \text{ sand: } P\left\{N(0,1) \geq \frac{\log 0.5617 + 6.5}{\sqrt{13}}\right\} \simeq 0.050 .$$

Den diskriminerende linie bliver nu bestemt ved

$$5x_1 - 2x_2 - 9\tfrac{1}{2} = \log 0.5617 ,$$

eller

$$5x_1 - 2x_2 - 8.92 = 0 .$$

Denne linie skærer forbindelseslinien mellem $\underline{\mu}_1$ og $\underline{\mu}_2$ i (2.36, 1.46), d.v.s. rykket mod $\underline{\mu}_2$ i forhold til midtpunktet (2.5, 1.5). Det er også umiddelbart klart, at linien parallelforskydes i denne retning; thi af tabsmatricen fremgår, at det er alvorligere at tage fejl, hvis $\underline{\mu}_1$ er sand, end hvis $\underline{\mu}_2$ er det. Vi skal derfor gøre R_1 større, i.e. rykke den begrænsende linie mod $\underline{\mu}_2$.

□

Vi må her præcisere, at det er afgørende, at dispersionsmatricerne for de to populationer er ens. Hvis dette ikke er tilfældet, får vi et helt andet resultat frem, som det vil fremgå af følgende eksempel.

Eksempel 7.3 Lad os nu antage, at dispersionsmatricen for population 2 er ændret til en enhedsmatrix, i.e.

$$\pi_1 \leftrightarrow N\left(\begin{pmatrix}4\\2\end{pmatrix}, \begin{pmatrix}1 & 1\\1 & 2\end{pmatrix}\right)$$

$$\pi_2 \leftrightarrow N\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\right) .$$

Vi vil igen søge at klassificere en observation $\underline{X}$, der følger en af ovenstående fordelinger. Da dispersionsmatricerne ikke er ens, kan vi ikke bruge resultatet i sætning 7.3, men vi må gå i gang fra bunden med sætning 7.2.

For $c > 0$ har vi

$$\frac{f_1(\underline{x})}{f_2(\underline{x})} \geq c \Leftrightarrow -(\underline{x}-\underline{\mu}_1)'\underline{\underline{\Sigma}}_1^{-1}(\underline{x}-\underline{\mu}_1)+(\underline{x}-\underline{\mu}_2)'\underline{\underline{\Sigma}}_2^{-1}(\underline{x}-\underline{\mu}_2) \geq 2\log c .$$

Da

$$\begin{aligned}(\underline{x}-\underline{\mu}_1)'\underline{\underline{\Sigma}}_1^{-1}(\underline{x}-\underline{\mu}_1) &= 2(x_1-4)^2+(x_2-2)^2-2(x_1-4)(x_2-2)\\ &= 2x_1^2+x_2^2-2x_1x_2-12x_1+4x_2+20 ,\end{aligned}$$

og

$$\begin{aligned}(\underline{x}-\underline{\mu}_2)'\underline{\underline{\Sigma}}_2^{-1}(\underline{x}-\underline{\mu}_2) &= (x_1-1)^2+(x_2-1)^2\\ &= x_1^2+x_2^2-2x_1-2x_2+2 ,\end{aligned}$$

er

7.14

$$\frac{f_1(\underline{x})}{f_2(\underline{x})} \geq c \Leftrightarrow -x_1^2+2x_1x_2+10x_1-6x_2-18 \geq 2 \log c \ .$$

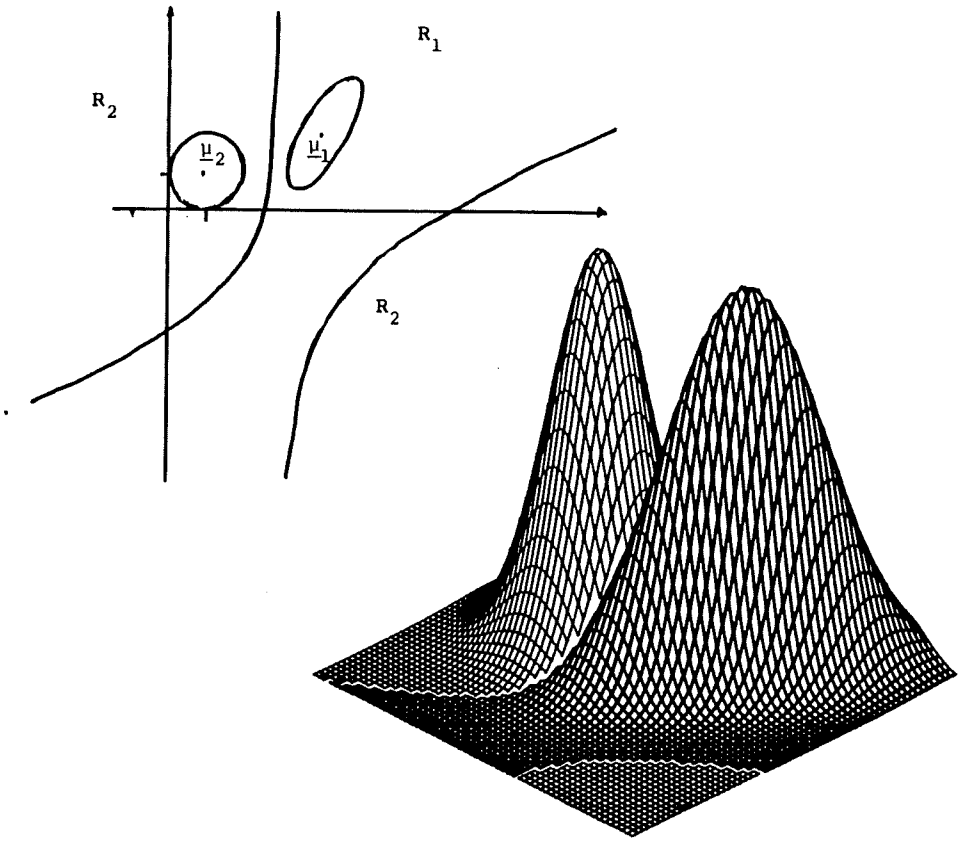
Vælger vi $c = 1$, ser vi, at kurven, der adskiller R_1 og R_2, er hyperblen

$$\{ \underline{x} \mid -x_1^2+2x_1x_2+10x_1-6x_2-18 = 0 \} \ .$$

Den har centrum i (3, -2) og asymptoterne

$$x_1 - 3 = 0 \ ,$$

$$x_1 - 2x_2 - 7 = 0 \ .$$



Disse kurver er sammen med konturellipserne for de to normale fordelinger indtegnet i ovenstående figur. Bemærk f. eks. at et punkt som (9,0) ligger i R_2 og altså bliver klassificeret som kommende fra fordelingen med centrum i (1,1). Endvidere er angivet selve frekvensfunktionerne. □

Vi skal ikke komme ind på problemet med misklassifikationssandsynligheder i tilfælde som ovenstående, hvor vi har kvadratiske diskriminatorer.

7.1.3 Diskrimination med ukendte parametre

Hvis man ikke kender de to fordelinger f_1 og f_2, må man estimere dem på basis af nogle observationer, og dernæst kan man så konstruere diskriminatorer ud fra de estimerede fordelinger, ganske som vi har gjort ud fra de eksakte.

Lad os betragte det normale tilfælde

$$\pi_1 \leftrightarrow N(\underline{\mu}_1, \underline{\underline{\Sigma}})$$

$$\pi_2 \leftrightarrow N(\underline{\mu}_2, \underline{\underline{\Sigma}}) \ ,$$

hvor parametrene er ukendte. Hvis vi har observationer $\underline{X}_1, \cdots, \underline{X}_{n_1}$, som vi <u>ved</u> stammer fra π_1, og observationer $\underline{Y}_1, \cdots, \underline{Y}_{n_2}$, som vi <u>ved</u> stammer fra π_2, kan vi estimere parametrene som følger

$$\hat{\underline{\mu}}_1 = \frac{1}{n_1} \sum_i \underline{X}_i = \overline{\underline{X}}$$

$$\hat{\underline{\mu}}_2 = \frac{1}{n_2} \sum_i \underline{Y}_i = \overline{\underline{Y}}$$

$$\hat{\underline{\underline{\Sigma}}} = \frac{1}{n_1+n_2-2} \left(\sum_i (\underline{X}_i - \overline{\underline{X}})(\underline{X}_i - \overline{\underline{X}})' + \sum_i (\underline{Y}_i - \overline{\underline{Y}})(\underline{Y}_i - \overline{\underline{Y}})' \right)$$

Vi har nu i fuldstændig analogi med sætningen p. 7.5 diskriminatoren

$$\underline{x}'\hat{\underline{\Sigma}}^{-1}(\hat{\underline{\mu}}_1-\hat{\underline{\mu}}_2) - \tfrac{1}{2}\hat{\underline{\mu}}_1'\hat{\underline{\Sigma}}^{-1}\hat{\underline{\mu}}_1 + \tfrac{1}{2}\hat{\underline{\mu}}_2'\hat{\underline{\Sigma}}^{-1}\hat{\underline{\mu}}_2$$

Den eksakte fordeling af denne størrelse, hvis vi erstatter $\underline{x}$ med en stokastisk variabel $\underline{X} \in N(\underline{\mu}_1, \underline{\Sigma})$, er ret kompliceret; men for store stikprøvestørrelser er den asymptotisk lig fordelingen af Z i sætning 7.5, således at vi for rimelige stikprøvestørrelser kan anvende den teori, vi har udledt.

Den estimerede norm mellem forventningsværdierne

$$||\hat{\underline{\mu}}_1-\hat{\underline{\mu}}_2||^2 \sim D^2 = (\hat{\underline{\mu}}_1-\hat{\underline{\mu}}_2)'\hat{\underline{\Sigma}}^{-1}(\hat{\underline{\mu}}_1-\hat{\underline{\mu}}_2) = ||\hat{\underline{\mu}}_1-\hat{\underline{\mu}}_2||^2_{\hat{\underline{\Sigma}}^{-1}} \cdot$$

kaldes Mahalanobis' afstand. Det må her indskydes, at en mængde forfattere benytter vendingen Mahalanobis' afstand også om størrelsen $||\underline{\mu}_1-\underline{\mu}_2||^2$ efter den indiske statistiker P.C. Mahalanobis, der samtidigt med men uafhængigt af den engelske statistiker R.A. Fisher udviklede diskriminantanalysen i trediverne.

Ved hjælp af D^2 kan vi i øvrigt teste, om $\underline{\mu}_1 = \underline{\mu}_2$, idet

$$Z = \frac{n_1+n_2-p-1}{p(n_1+n_2-2)} \cdot \frac{n_1 n_2}{n_1+n_2} D^2$$

er $F(p, n_1+n_2-p-1)$-fordelt, hvis $\underline{\mu}_1 = \underline{\mu}_2$. Hvis $\underline{\mu}_1 \neq \underline{\mu}_2$, har Z en større middelværdi, således at det kritiske område bliver store værdier af Z. Dette test er selvsagt ækvivalent med det i afsnit 6.1.2 anførte Hotellings T^2-test.

Vi anfører et eksempel (data stammer fra K.R. Nair: A biometric study of the desert locust, Bull. Int. Stat. Inst. 1951).

Eksempel 7.4 Ved en undersøgelse af ørkengræshopper har man målt forskellige biometriske karakteristika, nemlig

x_1: længde af bageste femur

x_2: maksimal bredde af hovedet i den genale region

x_3: længde af pronotum ved skallen.

De to arter, der er undersøgt, er gregaria og en mellemfase mellem gregaria og solotaria.

Man har fundet følgende middelværdier

	Middelværdier	
	Gregaria $n_1 = 20$	Mellemfase $n_2 = 72$
x_1	25.80	28.35
x_2	7.81	7.41
x_3	10.77	10.75

Den estimerede dispersionsmatrix er

	x_1	x_2	x_3
x_1	4.7350	0.5622	1.4685
x_2	0.5622	0.1413	0.2174
x_3	1.4685	0.2174	0.5702

Man er nu interesseret i at få opstillet en diskriminantfunktion til klassifikation af fremtidige græshopper ved hjælp af måleværdier af x_1, x_2, x_3.

Først må det dog være rimeligt at undersøge, om de 3 egenskaber overhovedet er forskellige for de to populationer, i.e. vi må undersøge, om det kan antages, at $\underline{\mu}_1 = \underline{\mu}_2$. Vi har

$$D^2 = (\hat{\underline{\mu}}_1 - \hat{\underline{\mu}}_2)' \hat{\underline{\underline{\Sigma}}}^{-1} (\hat{\underline{\mu}}_1 - \hat{\underline{\mu}}_2) = 9.7421 \ .$$

Denne værdi indsætter vi i teststørrelsen p. 7.16 og får

$$Z = \frac{20+72-3-1}{3(20+72-2)} \cdot \frac{20 \cdot 72}{20+72} \cdot 9.7421 = 49.70 \ .$$

Da

$$F(3,88)_{0.999} \simeq 6 \ ,$$

vil vi forkaste hypotesen, at de to middelværdier er ens, og det er derfor ikke urimeligt at forsøge at opstille en diskriminator.

Vi har

$$\underline{x}' \hat{\underline{\underline{\Sigma}}}^{-1} (\hat{\underline{\mu}}_1 - \hat{\underline{\mu}}_2) = - 2.7458 \, x_1 + 6.6217 \, x_2 + 4.5820 \, x_3$$

og

$$\frac{1}{2} (\hat{\underline{\mu}}_1' \hat{\underline{\underline{\Sigma}}}^{-1} \hat{\underline{\mu}}_1 - \hat{\underline{\mu}}_2' \hat{\underline{\underline{\Sigma}}}^{-1} \hat{\underline{\mu}}_2) = 25.3506 \ .$$

Da der ikke foreligger oplysninger om a priori sandsynligheder, vil vi anvende $c = 1$, d.v.s.: $\log c = 0$, og vi anvender altså funktionen

$$d(\underline{x}) = - 2.7458 \, x_1 + 6.6217 \, x_2 + 4.5820 \, x_3 - 25.3506$$

ved klassifikation af de to mulige græshoppearter.

Har vi eksempelvis fanget et eksemplar med de målte karakteristika

$$\underline{x} = \begin{pmatrix} 27.06 \\ 8.03 \\ 11.36 \end{pmatrix}$$

får vi $d(\underline{x}) = 5.5715 > 0$, hvorfor vi klassificerer eksemplaret som værende en gregaria.

□

7.1.4 Test for bedste diskriminantfunktion

Vi minder om, at den bedste diskriminator

$$\hat{\underline{\delta}} = \hat{\underline{\underline{\Sigma}}}^{-1}(\hat{\underline{\mu}}_1 - \hat{\underline{\mu}}_2) \ ,$$

kan fås ved maksimalisering af funktionen

$$\hat{\varphi}(\underline{d}) = \frac{[(\hat{\underline{\mu}}_1 - \hat{\underline{\mu}}_2)'\underline{d}]^2}{\underline{d}'\hat{\underline{\underline{\Sigma}}}\underline{d}} \ .$$

Maksimalværdien er

$$\hat{\varphi}(\hat{\underline{\delta}}) = \frac{[(\hat{\underline{\mu}}_1 - \hat{\underline{\mu}}_2)'\hat{\underline{\underline{\Sigma}}}^{-1}(\hat{\underline{\mu}}_1 - \hat{\underline{\mu}}_2)]^2}{(\hat{\underline{\mu}}_1 - \hat{\underline{\mu}}_2)'\hat{\underline{\underline{\Sigma}}}^{-1}(\hat{\underline{\mu}}_1 - \hat{\underline{\mu}}_2)} = D^2 \ ,$$

d.v.s. Mahalanobis' D^2 er den maksimale værdi af $\hat{\varphi}(\underline{d})$. For et vilkårligt (fast) $\underline{d}$ sætter vi nu

$$D_1^2 = \hat{\varphi}(\underline{d}) = \frac{[(\hat{\underline{\mu}}_1 - \hat{\underline{\mu}}_2)'\underline{d}]^2}{\underline{d}'\hat{\underline{\underline{\Sigma}}}\underline{d}} \ .$$

Vi kan da teste hypotesen, at den ved $\underline{d}$ bestemte lineære afbildning er den bedste diskriminator ved hjælp af teststørrelsen

$$Z = \frac{n_1+n_2-p-1}{p-1} \cdot \frac{n_1 n_2 (D^2 - D_1^2)}{(n_1+n_2)(n_1+n_2-2)+n_1 n_2 D_1^2} \ ,$$

der er $F(p-1,\ n_1+n_2-p-1)$-fordelt under hypotesen. Store værdier af Z er kritiske.

Vi skal ikke komme ind på begrundelsen for, at 0-hypotesefordelingen er, som den er, men blot konstatere, at Z giver et mål for, hvor meget "afstanden" mellem de to populationer er formindsket ved anvendelse af $\underline{d}$ i stedet for $\hat{\underline{\delta}}$. Hvis denne formind-

skelse er for stor, i.e. hvis Z er stor, vil vi ikke kunne antage, at $\underline{d}$ yder en lige så god skelnen mellem de to populationer som $\hat{\underline{\delta}}$.

Eksempel 7.5 I nedenstående tabel er der anført gennemsnit af 50 målinger af forskellige karakteristika på to forskellige Irisarter, nemlig Iris versicolor og Iris setosa. (Data stammer fra Fisher's undersøgelser 1936).

	Versicolor	Setosa	Differens
Bægerblads længde	5.936	5.006	0.930
Bægerblads bredde	2.770	3.428	-0.658
Kronblads længde	4.260	1.462	2.789
Kronblads bredde	1.326	0.246	1.080

Den estimerede dispersionsmatrix (baseret på 98 frihedsgrader) er

$$\hat{\underline{\underline{\Sigma}}} = \begin{bmatrix} 0.19534 & 0.09220 & 0.099626 & 0.03306 \\ & 0.12108 & 0.04718 & 0.02525 \\ & & 0.12549 & 0.039586 \\ & & & 0.02511 \end{bmatrix}$$

Heraf findes umiddelbart

$$\hat{\underline{\delta}} = \hat{\underline{\underline{\Sigma}}}^{-1}(\hat{\underline{\mu}}_1 - \hat{\underline{\mu}}_2) = \begin{bmatrix} -3.0692 \\ -18.0006 \\ 21.7641 \\ 30.7549 \end{bmatrix} .$$

Mahalanobis' afstand mellem middelværdierne er

$$D^2 = [0.930, -0.658, 2.789, 1.080] \begin{bmatrix} -3.0692 \\ -18.0006 \\ 21.7641 \\ 30.7549 \end{bmatrix} = 103.2119 .$$

Vi tester først, om det kan antages, at $\underline{\mu}_1 = \underline{\mu}_2$. Teststørrelsen bliver

$$\frac{50+50-4-1}{4(50+50-2)} \frac{50 \cdot 50}{50+50} \cdot 103.2119 = 625.3256$$

$$> F(4,95)_{0.9995} \simeq 5.5 \ .$$

Det vil ikke være rimeligt at antage $\underline{\mu}_1 = \underline{\mu}_2$.

Ved at se på differenserne mellem komponenterne i $\underline{\mu}_1$ og $\underline{\mu}_2$, ser vi, at tallet for versicolor er størst undtagen for x_2 (bægerbladets bredde). Da vi søger en lineær afbildning, der antager en "stor" værdi på $\underline{\mu}_1 - \underline{\mu}_2$, kunne man prøve med afbildningen

$$\underline{x}' \underline{d}_0 = x_1 - x_2 + x_3 + x_4 \ ,$$

hvor $\underline{d}_0$ altså er vektoren $\begin{bmatrix} 1 \\ -1 \\ 1 \\ 1 \end{bmatrix}$.

Vi vil teste, om det kan antages, at den bedste diskriminator har formen

$$\underline{\delta} = \text{konstant} \cdot \begin{bmatrix} 1 \\ -1 \\ 1 \\ 1 \end{bmatrix} = \text{konstant} \cdot \underline{d}_0 \ .$$

Vi bestemmer den til $\underline{d}_0$ svarende værdi af φ:

$$\frac{[(\hat{\underline{\mu}}_1 - \hat{\underline{\mu}}_2)' \underline{d}_0]^2}{\underline{d}_0' \hat{\underline{\underline{\Sigma}}} \underline{d}_0} = 61.9479 \ .$$

Teststørrelsen bliver

$$\frac{50+50-4-1}{4-1} \cdot \frac{50 \cdot 50 \ (103.2119-61.9479)}{(50+50)(50+50-2)+50 \cdot 50 \cdot 61.9479}$$

$$= 1984 > F(3,95)_{0.9995} \simeq 6.5 \ .$$

Vi må altså forkaste hypotesen og konstatere, at vi ikke kan antage, at den bedste diskriminator er af formen $x_1 - x_2 + x_3 + x_4$.

□

7.1.5 Test for yderligere information

Når man har fået forelagt målinger af en række variable på nogle individer med henblik på bestemmelse af en diskriminantfunktion, rejser der sig naturligt det spørgsmål, om det virkelig har været og er nødvendigt med alle målinger, eller om man kan nøjes med færre variable til at skille populationerne fra hinanden. Man kunne eksempelvis forestille sig, at det ville være tilstrækkeligt at måle <u>længden</u> af bægerblade og kronblade for at skelne mellem Iris versicolor og Iris setosa.

Vi formulerer disse overvejelser lidt mere præcist. Ved diskriminationen måler vi de variable $X_1, \cdots, X_p$. Vi vil opstille et test for at undersøge, om det kan antages, at de sidste q variable er overflødige ved diskriminationen.

Vi regner fremdeles med, at der foreligger n_1 observationer fra population π_1 og n_2 observationer fra population π_2. Vi sætter

$$\begin{bmatrix} X_1 \\ \cdot \\ \cdot \\ \cdot \\ X_{p-q} \end{bmatrix} = \underline{X}_1 \quad \text{og} \quad \begin{bmatrix} X_{p-q+1} \\ \cdot \\ \cdot \\ \cdot \\ X_p \end{bmatrix} = \underline{X}_2 \ ,$$

og foretager samme opspaltning af middelværdivektor og dispersionsmatrix

$$\underline{\mu}_i = \begin{bmatrix} \underline{\mu}_i^{(1)} \\ \underline{\mu}_i^{(2)} \end{bmatrix}$$

$$\underline{\underline{\Sigma}} = \begin{bmatrix} \underline{\underline{\Sigma}}_{11} & \underline{\underline{\Sigma}}_{12} \\ \underline{\underline{\Sigma}}_{21} & \underline{\underline{\Sigma}}_{22} \end{bmatrix} .$$

Vi beregner nu Mahalanobis' afstand mellem populationerne, dels på grundlag af den fulde information, i.e. samtlige p variable, og dels på grundlag af den reducerede information, i.e. de første p-q variable. Vi har altså

$$D_p^2 = (\hat{\underline{\mu}}_1 - \hat{\underline{\mu}}_2)' \hat{\underline{\underline{\Sigma}}}^{-1} (\hat{\underline{\mu}}_1 - \hat{\underline{\mu}}_2)$$

og

$$D_{p-q}^2 = (\underline{\mu}_1^{(1)} - \underline{\mu}_2^{(1)})' \hat{\underline{\underline{\Sigma}}}_{11}^{-1} (\underline{\mu}_1^{(1)} - \underline{\mu}_2^{(1)}) \ .$$

Et test for hypotesen, at de sidste q variable ej bidrager til øgning af diskriminationen, baseres på

$$Z = \frac{n_1+n_2-p-1}{q} \frac{n_1 n_2 (D_p^2 - D_{p-q}^2)}{(n_1+n_2)(n_1+n_2-2)+n_1 n_2 D_{p-q}^2} \ .$$

Det kan vises, at $Z \in F(q, n_1+n_2-p-1)$, hvis H_0 er sand. Vi forbigår beviset, men skal blot påpege, at Z "måler" den relative øgning i "afstanden" mellem populationerne, når vi går fra p-q variable til p variable. Det er derfor også intuitivt rimeligt, at vi forkaster hypotesen, at det er tilstrækkeligt med p-q variable, hvis Z er stor.

Vi giver nu et illustrativt

Eksempel 7.6 Vi vil undersøge, om det er tilstrækkeligt at måle længden af bæger- og kronblade for at skelne mellem de i eksempel 7.5 anførte Irisarter.

Vi udfører nu en helt sædvanlig diskriminantanalyse på de anførte data, men vi ser bort fra breddemålingerne. Den resulterende Mahalanobisafstand er

$$D_2^2 = 76.7082 \ ,$$

hvorfor teststørrelsen for den anførte hypotese bliver

$$\frac{50+50-4-1}{2}\,\frac{50\cdot 50\ (103.2119-76.7082)}{(50+50)(50+50-2)+50\cdot 50\cdot 76.7082}$$

$$= 15.6132 > F(2,95)_{0.9995} \simeq 8.25 \ .$$

Vi må derfor gå ud fra, at der i breddemålingerne indeholdes yderligere information, som kan tjene til at skelne setosa fra versicolor.

□

7.2 Diskrimination mellem flere populationer

7.2.1 Bayesløsning

Hovedideen i generaliseringen i dette afsnit er faktisk, at man sammenligner populationerne 2 og 2 som i de indledende afsnit og så til sidst vælger den mest "sandsynlige" population.

Vi betragter populationerne

$$\pi_1, \cdots, \pi_k ,$$

og på basis af målinger af p egenskaber (eller variable) ved et individ ønsker vi at klassificere dette som hørende til en af populationerne $\pi_1, \cdots, \pi_k$.

Måleresultatet er

$$\underline{X} = \begin{pmatrix} X_1 \\ \cdot \\ \cdot \\ \cdot \\ X_p \end{pmatrix} .$$

Hvis individet stammer fra π_i, er frekvensfunktionen for $\underline{X}$ $f_i(\underline{x})$.

Vi antager, at der er givet en tabsfunktion L, der er anført i nedenstående tabel.

		Vælger			
		π_1	π_2	$\cdots$	π_k
Tilstand	π_1	0	L(1,2)	$\cdots$	L(1,k)
	π_2	L(2,1)	0	$\cdots$	L(2,k)
	π_k	L(k,1)	L(k,2)	$\cdots$	0

Endelig kan vi antage, at der er en a priorifordeling

$$g(\pi_i) = p_i, \quad i = 1, \cdots, k .$$

For et individ med målingen $\underline{x}$ defineres dets diskriminantværdi (eng.: discriminant score) for den i'te population som

$$S_i^*(\underline{x}) = S_i^* = -[p_1 f_1(\underline{x})L(1,i) + \cdots + p_k f_k(\underline{x})L(k,i)]$$

(bemærk, at L(i,i) = 0, hvorfor der i summen ikke optræder et led med $p_i f_i(\underline{x})$). Da a posteriori-sandsynligheden for π_ν er

$$k(\pi_\nu|\underline{x}) = \frac{p_\nu f_\nu(\underline{x})}{p_1 f_1(\underline{x}) + \cdots + p_k f_k(\underline{x})}$$

$$= \frac{p_\nu f_\nu(\underline{x})}{h(\underline{x})} ,$$

ser vi, at S_i^* er en konstant ($-h(\underline{x})$) ganget det forventede tab med hensyn til a posteriori-fordelingen af π ved at vælge den i'te population. Da proportionalitetsfaktoren $-h(\underline{x})$ er negativ, ser vi, at Bayesløsningen til beslutningsproblemet er at vælge den population, der har den største diskriminantværdi, i.e. vælge π_ν, hvis

$$S_\nu^* \geq S_i^* \quad , \quad \forall i .$$

Hvis alle L(i,j) ($i \neq j$) er ens, kan vi simplificere udtrykket for diskriminantværdien. Vi foretrækker π_i frem for π_j, hvis

$$S_i^* > S_j^* \ ,$$

d.v.s. hvis

$$-(\sum_\nu p_\nu f_\nu(\underline{x}) - p_i f_i(\underline{x})) > -(\sum_\nu p_\nu f_\nu(\underline{x}) - p_j f_j(\underline{x}))$$

$$\leftrightarrow p_i f_i(\underline{x}) > p_j f_j(\underline{x}) \ .$$

Vi kan derfor i dette tilfælde vælge diskriminantværdien

$$S_i' = p_i f_i(\underline{x}) \ .$$

Bayesreglen er altså her, at vi vælger den population, der har den største a posteriori-sandsynlighed, ɔ: vælger gruppe i, hvis $S_i' > S_j'$, $\forall j \neq i$. Denne regel vælges ikke alene, hvor tabene er ens, men også hvor det ikke er muligt at fastsætte sådanne. Hvis p_i'erne ikke kendes, eller det ikke er muligt at estimere dem, vælges som regel diskriminantværdien

$$S_i'' = f_i(\underline{x}) \ ,$$

d.v.s. vi vælger den population, hvor den observerede sandsynlighed er størst.

Minimaxløsningerne bestemmes ved at vælge den strategi, der bevirker, at alle misklassifikationssandsynligheder bliver lige store. (Stadig under den forudsætning at alle tab er ens). Vi skal dog ikke nærmere komme ind på disse problemer.

7.2.2 Bayesløsning i tilfældet med flere normale fordelinger

Vi vil nu betragte det tilfælde, hvor

$$\pi_i \leftrightarrow N(\underline{\mu}_i, \underline{\underline{\Sigma}}_i) \ ,$$

d.v.s

$$f_i(\underline{x}) = \frac{1}{\sqrt{2\pi}^p} \frac{1}{\sqrt{\det \underline{\underline{\Sigma}}_i}} \exp(-\frac{1}{2}(\underline{x}-\underline{\mu}_i)' \underline{\underline{\Sigma}}_i^{-1} (\underline{x}-\underline{\mu}_i)) \ ,$$

for $i = 1, \cdots, k$.

Da vi får samme beslutningsregel ved at vælge monotone transformationer af vores diskriminantværdier, tager vi logaritmen af f_i'erne og ser bort fra den fælles faktor $(2\pi)^{-p/2}$. Dette giver (idet vi forudsætter, at tabene er ens)

$$S_i' = -\frac{1}{2}\log(\det\underline{\underline{\Sigma}}_i) - \frac{1}{2}(\underline{x}-\underline{\mu}_i)'\underline{\underline{\Sigma}}_i^{-1}(\underline{x}-\underline{\mu}_i) + \log p_i \ .$$

Denne funktion er kvadratisk i $\underline{x}$ og kaldes en kvadratisk diskriminantfunktion. Hvis alle $\underline{\underline{\Sigma}}_i$ er ens, er leddene

$$-\frac{1}{2}\log\det\underline{\underline{\Sigma}} - \frac{1}{2}\underline{x}'\underline{\underline{\Sigma}}^{-1}\underline{x} \ .$$

fælles for alle S_i'er og kan derfor udelades. Vi får da

$$S_i = \underline{x}'\underline{\underline{\Sigma}}^{-1}\underline{\mu}_i - \frac{1}{2}\underline{\mu}_i'\underline{\underline{\Sigma}}^{-1}\underline{\mu}_i + \log p_i \ .$$

Dette er åbenbart en lineær (affin) funktion i $\underline{x}$. Hvis der kun er 2 grupper, ser vi, at vi vælger gruppe 1, netop hvis

$$S_1' > S_2' \Leftrightarrow S_1 - S_2 > 0$$

$$\Leftrightarrow \underline{x}'\underline{\underline{\Sigma}}^{-1}(\underline{\mu}_1-\underline{\mu}_2) - \frac{1}{2}\underline{\mu}_1'\underline{\underline{\Sigma}}^{-1}\underline{\mu}_1 + \frac{1}{2}\underline{\mu}_2'\underline{\underline{\Sigma}}^{-1}\underline{\mu}_2 \geq \log\frac{p_2}{p_1} \ ,$$

d.v.s. det samme resultat som p. 7.5.

A posteriori-sandsynligheden for den ν'te gruppe bliver

$$k(\pi_\nu|\underline{x}) = \frac{\exp(S_\nu)}{\sum_{i=1}^{k}\exp(S_i)}$$

Det er naturligvis muligt at beskrive beslutningsreglerne ved en inddeling af R^p i mængder $R_1, \cdots, R_k$, således at vi vælger π_i, netop når $\underline{x} \in R_i$. Det vil bl.a. fremgå af følgende

Eksempel 7.7 Vi betragter populationer π_1, π_2 og π_3 givet ved normale fordelinger med forventningsværdier

$$\underline{\mu}_1 = \begin{pmatrix}4\\2\end{pmatrix}, \quad \underline{\mu}_2 = \begin{pmatrix}1\\1\end{pmatrix} \text{ og } \underline{\mu}_3 = \begin{pmatrix}2\\6\end{pmatrix},$$

og den fælles dispersionsmatrix

$$\underline{\underline{\Sigma}} = \begin{pmatrix}1 & 1\\1 & 2\end{pmatrix}$$

jvf. eksemplet p. 7.10. Vi har da - idet vi antager, at alle p_i er ens, hvorfor vi kan udelade dem fra diskriminantværdierne -

$$S'_{11} = (x_1x_2)\begin{pmatrix}2 & -1\\-1 & 1\end{pmatrix}\begin{pmatrix}4\\2\end{pmatrix} - \tfrac{1}{2}(4,2)\begin{pmatrix}2 & -1\\-1 & 1\end{pmatrix}\begin{pmatrix}4\\2\end{pmatrix}$$

$$= 6x_1 - 2x_2 - 10$$

$$S'_{12} = (x_1x_2)\begin{pmatrix}2 & -1\\-1 & 1\end{pmatrix}\begin{pmatrix}1\\1\end{pmatrix} - \tfrac{1}{2}(1,1)\begin{pmatrix}2 & -1\\-1 & 1\end{pmatrix}\begin{pmatrix}1\\1\end{pmatrix}$$

$$= x_1 - \tfrac{1}{2}$$

$$S'_{13} = (x_1x_2)\begin{pmatrix}2 & -1\\-1 & 1\end{pmatrix}\begin{pmatrix}2\\6\end{pmatrix} - \tfrac{1}{2}(2,6)\begin{pmatrix}2 & -1\\-1 & 1\end{pmatrix}\begin{pmatrix}2\\6\end{pmatrix}$$

$$= -2x_1 + 4x_2 - 10 .$$

Vi vælger nu π_1 frem for π_2, hvis

$$u_{12}(\underline{x}) = 6x_1 - 2x_2 - 10 - (x_1 - \tfrac{1}{2})$$

$$= 5x_1 - 2x_2 - 9\tfrac{1}{2}$$

$$> 0 .$$

Vi vælger π_1 frem for π_3, hvis

$$u_{13}(\underline{x}) = 6x_1 - 2x_2 - 10 - (-2x_1+4x_2-10)$$

$$= 8x_1 - 6x_2$$

$$> 0 \ ,$$

og endelig π_2 frem for π_3, hvis

$$u_{23}(\underline{x}) = x_1 - \tfrac{1}{2} - (-2x_1+4x_2-10)$$

$$= 3x_1 - 4x_2 + 9\tfrac{1}{2}$$

$$> 0 \ .$$

Det er nu evident, at vi vælger π_1, hvis såvel $u_{12}(\underline{x}) > 0$ som $u_{13}(\underline{x}) > 0$, og analogt med de øvrige.

Vi kan derfor definere områderne

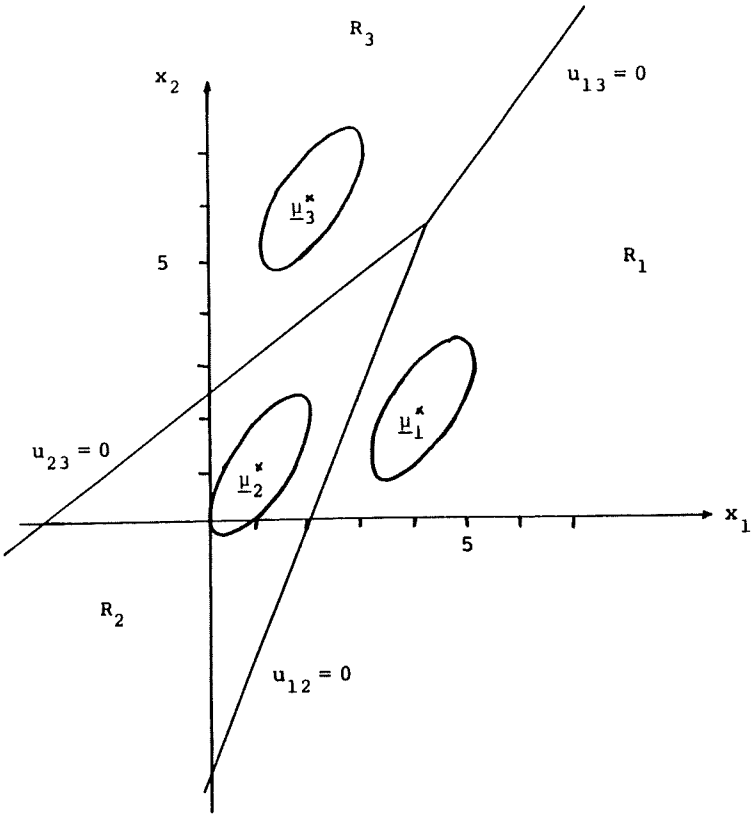
$$R_1 = \{\underline{x} \mid u_{12}(\underline{x}) > 0 \wedge u_{13}(\underline{x}) > 0\}$$

$$R_2 = \{\underline{x} \mid u_{12}(\underline{x}) < 0 \wedge u_{23}(\underline{x}) > 0\}$$

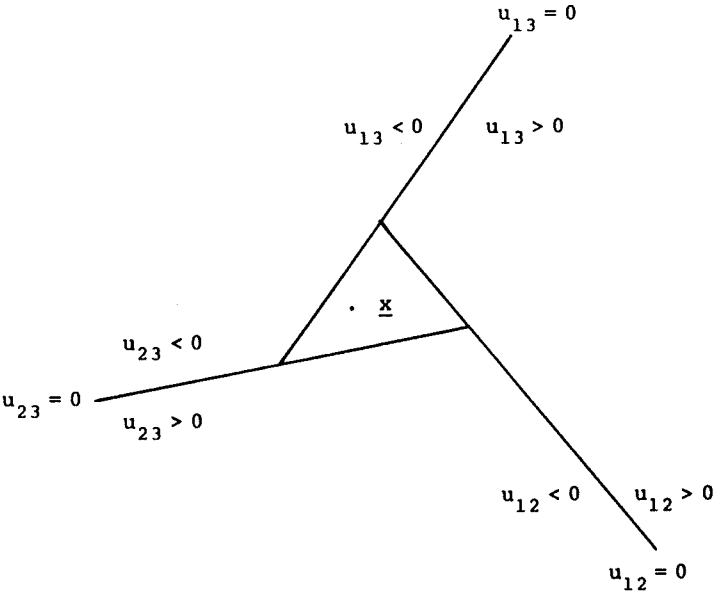
$$R_3 = \{\underline{x} \mid u_{13}(\underline{x}) < 0 \wedge u_{23}(\underline{x}) < 0\} \ ,$$

og vi har da, at vi vælger π_i, netop hvis $\underline{x} \in R_i$.

Vi har skitseret situationen i nedenstående tegning.



Man kan let direkte overtyde sig om, at linierne skærer hinanden i et punkt. Det er dog også muligt at ræsonnere sig frem til dette. Lad os antage, at situationen er som i nedenstående skitse



Vi bemærker nu, at

$$u_{ij} > 0 \Leftrightarrow S'_{1i} > S'_{1j} \Leftrightarrow f_i > f_j \ .$$

Om punktet $\underline{x}$ gælder

$$\left.\begin{array}{l} u_{23}(\underline{x}) < 0 \text{ d.v.s. } f_2(\underline{x}) < f_3(\underline{x}) \\ u_{13}(\underline{x}) > 0 \text{ d.v.s. } f_1(\underline{x}) > f_3(\underline{x}) \end{array}\right\} \Rightarrow f_1(\underline{x}) > f_2(\underline{x})$$

$$u_{12}(\underline{x}) < 0 \text{ d.v.s. } f_1(\underline{x}) < f_2(\underline{x})$$

d.v.s. vi har etableret en modstrid, d.v.s. de 3 linier bestemt ved u_{12}, u_{13} og u_{23} må skære hinanden i et punkt.

□

Hvis parametrene ikke er kendte, men estimerede, indsættes de estimerede udtryk i de ovenfor omtalte relationer, jvf. fremgangsmåden i afsnit 7.1.3.

7.2.3 Alternativ diskriminationsprocedure i tilfældet flere populationer

I det foregående afsnit har vi givet én form for generalisering af diskriminantanalysen fra to til flere populationer. Vi skal nu søge en anden fremgangsmåde, der i stedet generaliserer sætning 7.4.

Vi betragter fremdeles k grupper med $n_1, \cdots, n_k$ observationer i hver. Gruppegennemsnittene kaldes $\underline{\bar{X}}_1, \cdots, \underline{\bar{X}}_k$. Vi definerer en "mellem grupper" matrix (among groups)

$$\underline{\underline{A}} = \sum_{i=1}^{k} n_i (\underline{\bar{X}}_i - \underline{\bar{X}})(\underline{\bar{X}}_i - \underline{\bar{X}})' ,$$

en "inden for grupper" matrix (within groups)

$$\underline{\underline{W}} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\underline{X}_{ij} - \underline{\bar{X}}_i)(\underline{X}_{ij} - \underline{\bar{X}}_i)'$$

og en "total" matrix (total sum of squares)

$$\underline{\underline{T}} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\underline{X}_{ij} - \underline{\bar{X}})(\underline{X}_{ij} - \underline{\bar{X}})' .$$

En fundamental ligning er

$$\underline{\underline{T}} = \underline{\underline{A}} + \underline{\underline{W}} .$$

Vi kan nu gå i gang med selve diskriminationen. Vi søger en bedste diskriminatorfunktion, hvor "bedste" skal betyde, at funktionen skal maksimalisere forholdet mellem "variationen" mellem grupper og variationen inden for grupper, i.e. vi søger funktion $\underline{y} = \underline{d}'\underline{x}$, så

$$\varphi(\underline{d}) = \frac{\underline{d}'\underline{\underline{A}}\,\underline{d}}{\underline{d}'\underline{\underline{W}}\,\underline{d}} \qquad (\underline{d} \text{ vælges, så } \underline{d}'\underline{d} = 1)$$

maksimaliseres. Det følger af sætning 1.23, at maksimalværdien er den største egenværdi λ_1 og tilhørende egenvektor $\underline{d}_1$ til

$$\det(\underline{\underline{A}}-\lambda\underline{\underline{W}}) = 0$$

eller

$$\det(\underline{\underline{W}}^{-1}\underline{\underline{A}}-\lambda\underline{\underline{I}}) = 0 \ .$$

Vi søger dernæst en ny diskriminantfunktion $\underline{d}_2$, så

$$\varphi(\underline{d}_2) = \frac{\underline{d}_2\underline{\underline{A}}\,\underline{d}_2}{\underline{d}_2\underline{\underline{W}}\,\underline{d}_2}$$

maksimaliseres under bibetingelsen, at

$$\underline{d}_2'\,\underline{d}_1 = 0 \quad \text{eller} \quad \underline{d}_1 \perp \underline{d}_2 \quad \text{og} \quad \underline{d}_2'\,\underline{d}_2 = 1 \ .$$

Dette svarer til den næststørste egenværdi for $\underline{\underline{W}}^{-1}\underline{\underline{A}}$ og tilsvarende egenvektor.

Sådan kan fortsættes, indtil man når en egenværdi for $\underline{\underline{W}}^{-1}\underline{\underline{A}}$, der er 0 (eller indtil $\underline{\underline{W}}^{-1}\underline{\underline{A}}$ er fuldstændig udtømt).

Et plot af de enkelte observationers projektioner (normerede med total middel) ned på $\underline{d}_1,\underline{d}_2$ planen vil være meget nyttigt som visuelt hjælpemiddel. Det er denne plan, der separerer punkterne bedst i ovennævnte forstand.

Projektionernes koordinater bliver

$$[\underline{d}_1'(\underline{x}_{ij}-\bar{\underline{x}}) \ , \quad \underline{d}_2'(\underline{x}_{ij}-\bar{\underline{x}})] \ .$$

Et andet nyttigt plot består af vektorerne

$$\begin{pmatrix} d_{11} \\ d_{21} \end{pmatrix}, \ \ldots \begin{pmatrix} d_{1p} \\ d_{2p} \end{pmatrix} \ .$$

Disse angiver, med hvilken vægt værdien af de enkelte variable indgår i plottet på $(\underline{d}_1,\underline{d}_2)$-planen.

I f. eks. programmet BMD07M - STEPWISE DISCRIMINANT ANALYSIS - benævnes $(\underline{d}_1, \underline{d}_2)$ planen "the first two canonical variables".

I dette program kan variable - som navnet antyder - tages ind og ud af analyse på en måde, som er fuldstændig analog til en trinvis regressionsanalyse (den version, der kaldes STEPWISE REGRESSION). Foruden at styre ind- og udtagning af variable ved hjælp af F-tests findes en række andre intuitive kriterier, som er udmærket beskrevet i BMD-manualen p. 243.

Her bør også nævnes, at Wilks Λ for test af hypotesen

$$H_o: \underline{\mu}_1 = \cdots = \underline{\mu}_k \quad \text{mod} \quad H_1: \exists i,j: \underline{\mu}_i \neq \underline{\mu}_j \ ,$$

er

$$\Lambda = \frac{\det \underline{\underline{W}}}{\det \underline{\underline{T}}} = \prod_{j=1}^{p} \frac{1}{1+\lambda_j} \ .$$

Denne størrelses fordeling kan approximeres ved en χ^2- eller F-fordeling. Den sidste mulighed er nok den numerisk bedste approximation. Disse er anført i BMD-manualen p. 242. Jævnfør i øvrigt med afsnit 6.3.1.

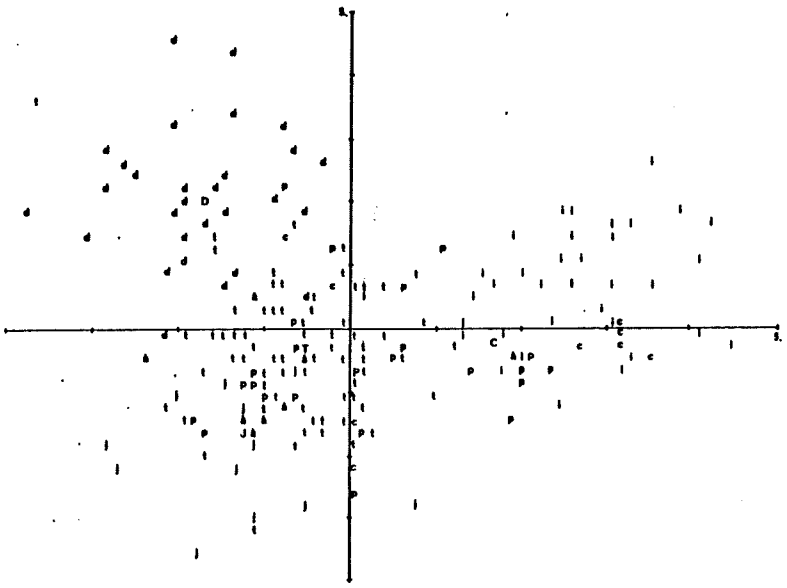
Exempel 7.8 I nedenstående tabel gives middelværdi og standardafvigelse af elementindhold for 208 vaskeprøver indsamlet i Jameson Land. Variablen Sum svarer til summen af Y og La indholdet.

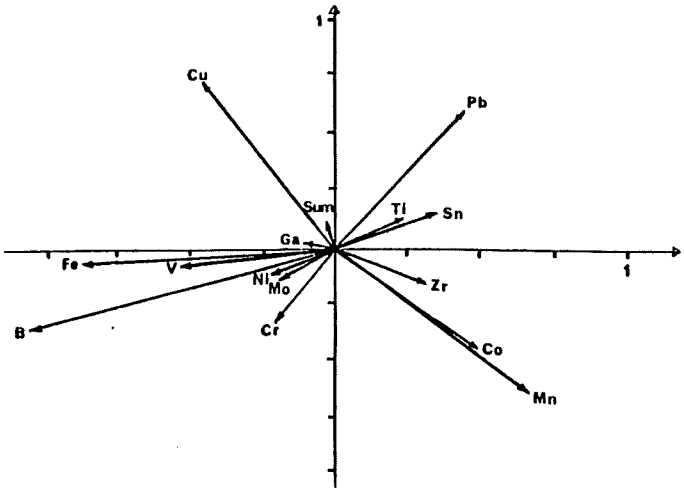
Variabel	Middel	Stand.afv.
B	73	141
Ti	40563	22279
V	678	491
Cr	1135	1216
Mn	2562	2081
Fe	225817	122302
Co	62	26
Ni	116	54
Cu	69	56
Ga	21	10
Zr	14752	14771
Mo	29	20
Sn	56	99
Pb	351	786
Sum	-	-

En fordelingsanalyse viste, at data bedst approximeredes ved LN-fordelinger. Derfor transformeredes alle tal, og de blev standardiseret for at opnå middel 0 og varians 1. Problemet er, i hvor høj grad elementindholdene karakteriserer de forskellige geologiske perioder. Antallet af målinger fra de enkelte perioder er anført nedenfor.

Periode	Antal
Jura	17
Trias	80
Perm	30
Kul	9
Devon	31
Tertiære intrusiver	35
Caledonsk krystallinsk	4
Eleonora Bay Formation	2

For at undersøge dette er udført nogle diskriminantanalyser. Dette skal vi ikke komme ind på her. Vi vil blot illustrere anvendelsen af det foran omtalte plot, der er anført nedenfor





I ovenstående figur er vist koefficienterne for de almindelige variable på de to "kanoniske" variable.

Ved at sammenligne de to figurer ser man bl.a., at Cu er temmelig specifik for Devon, og overhovedet giver billederne et godt indtryk af, hvorledes elementfordelingerne er for de forskellige perioder.

□

7.3 Nogle standardprogrammer til beregning af lineære diskriminatorer

Nogle af de mere anvendte diskriminantanalyseprogrammer er to programmer fra BMD-manualen (BMDO4M og BMDO5M) samt SSP sample programmet MDISC.

BMD programmerne behandler 2 henholdsvis 2 eller flere populationer. SSP programmet behandler 2 eller flere populationer.

Endvidere må nævnes det i foregående afsnit omtalte "stepwise" program fra BMD-pakken (BMDO7M).

SSP programmet er opbygget på samme måde som BMDO5M, hvad angår beregningsmetoder. Det har dog ikke direkte helt så mange muligheder for et varieret input. Det er dog ret enkelt at ændre i hovedprogrammet MDISC, hvis det skal tilpasses et specielt inputmateriale.

De fleste af de størrelser, der anføres i output, vil umiddelbart kunne forstås med 2 undtagelser, nemlig størrelsen "generalized Mahalanobis D-square" og tallene "probabilities associated with largest discriminant function". Vi skal nu kort omtale, hvad der menes med disse begreber.

Lad der være givet k populationer $\pi_1, \cdots, \pi_k$, og lad $\pi_i \leftrightarrow N_p(\underline{\mu}_i, \underline{\underline{\Sigma}})$. Vi forudsætter, at der foreligger observationer

$$\begin{aligned} &\underline{X}_{11}, \cdots, \underline{X}_{1n_1} \quad \text{fra } \pi_1 \\ &\quad \vdots \\ &\underline{X}_{k1}, \cdots, \underline{X}_{kn_k} \quad \text{fra } \pi_k \ . \end{aligned}$$

Ved den generaliserede Mahalanobis D^2-størrelse mellem populationerne $\pi_1, \cdots, \pi_k$ forstås størrelsen

$$V = \sum_{i=1}^{k} n_i (\underline{\bar{X}}_i - \underline{\bar{X}})' \hat{\underline{\underline{\Sigma}}}^{-1} (\underline{\bar{X}}_i - \underline{\bar{X}}) \ ,$$

hvor $\underline{\bar{X}}_i$ er gennemsnittet af målingerne fra i'te population.

Vi vil nu bestemme den approximative fordeling for V. Hvis $\underline{X} \in N_p(\underline{\mu}, \underline{\underline{\Sigma}})$, gælder ifølge sætning 2.15, at

$$||\underline{X}-\underline{\mu}||^2 = (\underline{X}-\underline{\mu})' \underline{\underline{\Sigma}}^{-1} (\underline{X}-\underline{\mu}) \in \chi^2(p) \ .$$

Af χ^2-fordelingens reproduktivitetssætning fås - såfremt alle $\underline{\mu}_i = \underline{\mu}$ -

$$\sum_{i=1}^{k} n_i (\underline{\bar{X}}_i - \underline{\mu})' \underline{\underline{\Sigma}}^{-1} (\underline{\bar{X}}_i - \underline{\mu}) \in \chi^2(kp) ,$$

og dermed, at

$$\sum_{i=1}^{k} n_i (\underline{\bar{X}}_i - \underline{\mu})' \hat{\underline{\underline{\Sigma}}}^{-1} (\underline{\bar{X}}_i - \underline{\mu}) \text{ approximativt } \in \chi^2(kp) .$$

Estimeres nu de p værdier i $\underline{\mu}$, fås stadig en approximativ χ^2-fordeling, men med p frihedsgrader færre, i.e.

$$V = \sum_{i=1}^{k} n_i (\underline{\bar{X}}_i - \underline{\bar{X}})' \hat{\underline{\underline{\Sigma}}}^{-1} (\underline{\bar{X}}_i - \underline{\bar{X}}) \text{ approximativt } \in \chi^2(p(k-1)),$$

såfremt alle $\underline{\mu}_i$ er ens. Størrelsen V kan derfor bruges som teststørrelse ved test af hypotesen

$$H_0 : \underline{\mu}_1 = \cdots = \underline{\mu}_k \text{ mod } H_1 : \exists\, i,j(\underline{\mu}_i \neq \underline{\mu}_j) .$$

Det kritiske område er naturligvis givet ved store værdier af V. Testet er dog et dårligere test end det i afsnit 6.3.1 anførte.

For k = 2 går $V = V_2$ ikke direkte over i Mahalanobis' afstand, men vi har

$$V_2 = \sum_{i=1}^{2} n_i (\underline{\bar{X}}_i - \underline{\bar{X}})' \hat{\underline{\underline{\Sigma}}}^{-1} (\underline{\bar{X}}_i - \underline{\bar{X}})$$

$$= \frac{n_1 n_2}{n_1 + n_2} (\underline{\bar{X}}_1 - \underline{\bar{X}}_2)' \hat{\underline{\underline{\Sigma}}}^{-1} (\underline{\bar{X}}_1 - \underline{\bar{X}}_2) ,$$

eller

$$V_2 = \frac{n_1 n_2}{n_1 + n_2} D^2 .$$

Kalder vi den i'te diskriminantfunktion (jvf. p. 7.27) g_i, d.v.s.

$$g_i(\underline{x}) = \underline{x}'\hat{\underline{\Sigma}}^{-1}\hat{\underline{\mu}}_i - \frac{1}{2}\hat{\underline{\mu}}_i'\hat{\underline{\Sigma}}^{-1}\hat{\underline{\mu}}_i \ ,$$

kan man for hver observation beregne størrelsen

$$p_{ij} = \frac{1}{\sum_{\nu=1}^{m} \exp(g_\nu(\underline{x}_{ij}) - c_{ij})} \ ,$$

hvor

$$c_{ij} = \max_{\mu} g_\mu(\underline{x}_{ij}) \ ,$$

d.v.s c_{ij} er den maksimale diskriminantværdi for den j'te observation fra population i. Det er klart, at $0 < p_{ij} < 1$.

Hvis p_{ij} er nær 1, betyder det, at

$$\max_{\mu} g_\mu(\underline{x}_{ij}) >> g_\nu(\underline{x}_{ij})$$

for alle ν bortset fra det ν, der giver maksimumsværdien. Dette betyder, at observationen ligger meget nær ved et enkelt $\hat{\underline{\mu}}_i$ og langt fra de øvrige.

Hvis p_{ij} er nær 0, er

$$\max_{\mu} g_\mu(\underline{x}_{ij}) \sim g_\nu(\underline{x}_{ij})$$

for alle ν, og det indebærer, at $\underline{x}_{ij}$ ligger nogenlunde lige langt fra alle $\hat{\underline{\mu}}_i$.

Det er nu klart, at man ved hjælp af disse størrelser p_{ij} kan danne sig et overblik over, hvor godt de estimerede diskriminantfunktioner klassificerer de foreliggende observationer. Det er p_{ij} størrelserne, der betegnes "probabilities" associated with largest discriminant function i forskellige standardprogrammer.

```
C       SAMPLE MAIN PROGRAM FOR DISCRIMINANT ANALYSIS - MDISC                  00070000
C                                                                              00080000
C       PURPOSE                                                                00090000
C          (1) READ THE PROBLEM PARAMETER CARD AND DATA FOR DISCRIMI-          00100000
C          NANT ANALYSIS, (2) CALL THREE SUBROUTINES TO CALCULATE VARI-        00110000
C          ABLE MEANS IN EACH GROUP, POOLED DISPERSION MATRIX, COMMON          00120000
C          MEANS OF VARIABLES, GENERALIZED MAHALANOBIS D SQUARE,               00130000
C          COEFFICIENTS OF DISCRIMINANT FUNCTIONS, AND PROBABILITY             00140000
C          ASSOCIATED WITH LARGEST DISCRIMINANT FUNCTION OF EACH               00150000
C          CASE IN EACH GROUP, AND (3) PRINT THE RESULTS.                      00160000
C                                                                              00170000
C       REMARKS                                                                00180000
C          THE NUMBER OF VARIABLES MUST BE GREATER THAN OR EQUAL TO            00190000
C          THE NUMBER OF GROUPS.                                               00200000
C                                                                              00210000
C       SUBROUTINES AND FUNCTION SUBPROGRAMS REQUIRED                          00220000
C          DMATX                                                               00230000
C          MINV                                                                00240000
C          DISCR                                                               00250000
C                                                                              00260000
C       METHOD                                                                 00270000
C          REFER TO '8MD COMPUTER PROGRAMS MANUAL', EDITED BY W. J.            00280000
C          DIXON, UCLA, 1964, AND T. W. ANDERSON, 'INTRODUCTION TO             00290000
C          MULTIVARIATE STATISTICAL ANALYSIS', JOHN WILEY AND SONS,            00300000
C          1958, SECTION 6.6-6.8.                                              00310000
C                                                                              00320000
C     ...................................................................      00330000
C                                                                              00340000
C     THE FOLLOWING DIMENSION MUST BE GREATER THAN OR EQUAL TO THE             00350000
C     NUMBER OF GROUPS, K..                                                    00360000
C                                                                              00370000
        DIMENSION N(5)                                                         00380000
C                                                                              00390000
C     THE FOLLOWING DIMENSION MUST BE GREATER THAN OR EQUAL TO THE             00400000
C     NUMBER OF VARIABLES, M..                                                 00410000
C                                                                              00420000
        DIMENSION CMEAN(10)                                                    00430000
C                                                                              00440000
C     THE FOLLOWING DIMENSION MUST BE GREATER THAN OR EQUAL TO THE             00450000
C     PRODUCT OF M*K..                                                         00460000
C                                                                              00470000
        DIMENSION XBAR(50)                                                     00480000
C                                                                              00490000
C     THE FOLLOWING DIMENSION MUST BE GREATER THAN OR EQUAL TO THE             00500000
C     PRODUCT OF (M+1)*K..                                                     00510000
C                                                                              00520000
        DIMENSION C(55)                                                        00530000
C                                                                              00540000
C     THE FOLLOWING DIMENSION MUST BE GREATER THAN OR EQUAL TO THE             00550000
C     PRODUCT OF M*M..                                                         00560000
C                                                                              00570000
        DIMENSION D(100)                                                       00580000
C                                                                              00590000
C     THE FOLLOWING DIMENSIONS MUST BE GREATER THAN OR EQUAL TO THE            00600000
C     TOTAL OF SAMPLE SIZES OF K GROUPS COMBINED, T (T = N(1)+N(2)+...         00610000
C     +N(K)).                                                                  00620000
C                                                                              00630000
        DIMENSION P(250),LG(250)                                               00640000
C                                                                              00650000
C     THE FOLLOWING DIMENSION MUST BE GREATER THAN OR EQUAL TO THE             00660000
C     TOTAL DATA POINTS WHICH IS EQUAL TO THE PRODUCT OF T*M..                  00670000
C                                                                              00680000
        DIMENSION X(2500)                                                      00690000
C                                                                              00700000
C     ...................................................................      00710000
C                                                                              00720000
C       IF A DOUBLE PRECISION VERSION OF THIS ROUTINE IS DESIRED, THE          00730000
C       C IN COLUMN 1 SHOULD BE REMOVED FROM THE DOUBLE PRECISION              00740000
C       STATEMENT WHICH FOLLOWS.                                               00750000
C                                                                              00760000
C     DOUBLE PRECISION CMEAN,XBAR,D,DET,C,V,P                                  00770000
C                                                                              00780000
C       THE C MUST ALSO BE REMOVED FROM DOUBLE PRECISION STATEMENTS            00790000
C       APPEARING IN OTHER ROUTINES USED IN CONJUNCTION WITH THIS              00800000
C       ROUTINE.                                                               00810000
C                                                                              00820000
C       ...................................................................    00830000
C                                                                              00840000
      1 FORMAT(4A4,A2,2I2,1215/(1415))                                         00850000
      2 FORMAT(27H1DISCRIMINANT ANALYSIS.....,A4,A2/19HO  NUMBER OF GROUPS=00860000
     1,I7X,13/22H   NUMBER OF VARIABLES=,I7/17H   SAMPLE SIZES=../12X,5HGRO00870000
     2UP)                                                                      00880000
      3 FORMAT(12X,I3,8X,I4)                                                   00890000
      4 FORMAT(1HO)                                                            00900000
      5 FORMAT(F4.0,4X,F4.0,4X)                                                00910000
      6 FORMAT(6HOGROUP,I3,7H  MEANS/(8F15.5))                                 00920000
      7 FORMAT(1HO/25H POOLED DISPERSION MATRIX)                               00930000
      8 FORMAT(4HOROW,I3/(8F15.5))                                             00940000
      9 FORMAT(1HO//13H COMMON MEANS/(8F15.5))                                 00950000
     10 FORMAT(1H///33H GENERALIZED MAHALANOBIS D-SQUARE=,F15.5//)             00960000
     11 FORMAT(22HODISCRIMINANT FUNCTION,I3/1H ,6X,27HCONSTANT    *   COEFF00970000
     1ICIENTS/1H F14.5,7H    *   ,7F14.5/(22X,7F14.5))                        00980000
     12 FORMAT(1HO//60H EVALUATION OF CLASSIFICATION FUNCTIONS FOR EACH OB00990000
     1SERVATION)                                                               01000000
     13 FORMAT(6HOGROUP,I3/19X,27HPROBABILITY ASSOCIATED WITH,11X,7HLARGEST01010000
     1T/13H  OBSERVATION,5X,29HLARGEST DISCRIMINANT FUNCTION,8X,12HFUNCTO1020000
     2ION NO.)                                                                 01030000
     14 FORMAT(1H ,I7,20X,F8.5,20X,I6)                                         01040000
C                                                                              01050000
```

```
C
CC      READ PROBLEM PARAMETER CARD                                          01060000
C                                                                            01070000
C                                                                            01080000
  100 READ (5,1) PR,PR1,K,M,(N(I),I=1,K)                                     01090000
C         PR.......PROBLEM NUMBER (MAY BE ALPHAMERIC)                        01100000
C         PR1......PROBLEM NUMBER (CONTINUED)                                01110000
CC        K........NUMBER OF GROUPS                                          01120000
CC        M........NUMBER OF VARIABLES                                       01130000
C         N........VECTOR OF LENGTH K CONTAINING SAMPLE SIZES                01140000
C                                                                            01150000
      WRITE (6,2) PR,PR1,K,M                                                 01160000
      DO 110 I=1,K                                                           01170000
  110 WRITE (6,3) I,N(I)                                                     01180000
      WRITE (6,4)                                                            01190000
C                                                                            01200000
CC    READ DATA                                                              01210000
C                                                                            01220000
      L=0                                                                    01230000
      DO 130 I=1,K                                                           01240000
      N1=N(I)                                                                01250000
      DO 120 J=1,N1                                                          01260000
      READ (5,5) (CMEAN(IJ),IJ=1,M)                                          01270000
      L=L+1                                                                  01280000
      N2=L-N1                                                                01290000
      DO 120 IJ=1,M                                                          01300000
      N2=N2+N1                                                               01310000
  120 X(N2)=CMEAN(IJ)                                                        01320000
  130 L=N2                                                                   01330000
C                                                                            01340000
      CALL DMATX (K,M,N,X,XBAR,D,CMEAN)                                      01350000
C                                                                            01360000
CC    PRINT MEANS AND POOLED DISPERSION MATRIX                               01370000
C                                                                            01380000
      L=0                                                                    01390000
      DO 150 I=1,K                                                           01400000
      DO 140 J=1,M                                                           01410000
      L=L+1                                                                  01420000
  140 CMEAN(J)=XBAR(L)                                                       01430000
  150 WRITE (6,6) I,(CMEAN(J),J=1,M)                                         01440000
      WRITE (6,7)                                                            01450000
      DO 170 I=1,M                                                           01460000
      L=I-M                                                                  01470000
      DO 160 J=1,M                                                           01480000
      L=L+M                                                                  01490000
  160 CMEAN(J)=D(L)                                                          01500000
  170 WRITE (6,8) I,(CMEAN(J),J=1,M)                                         01510000
C                                                                            01520000
      CALL MINV (D,M,DET,CMEAN,C)                                            01530000
C                                                                            01540000
      CALL DISCR (K,M,N,X,XBAR,D,CMEAN,V,C,P,LG)                             01550000
C                                                                            01560000
CC    PRINT COMMON MEANS                                                     01570000
C                                                                            01580000
      WRITE (6,9) (CMEAN(I),I=1,M)                                           01590000
C                                                                            01600000
CC    PRINT GENERALIZED MAHALANOBIS D-SQUARE                                 01610000
C                                                                            01620000
      WRITE (6,10) V                                                         01630000
C                                                                            01640000
CC    PRINT CONSTANTS AND COEFFICIENTS OF DISCRIMINANT FUNCTIONS             01650000
C                                                                            01660000
      N1=1                                                                   01670000
      N2=M+1                                                                 01680000
      DO 180 I=1,K                                                           01690000
      WRITE (6,11) I,(C(J),J=N1,N2)                                          01700000
      N1=N1+(M+1)                                                            01710000
  180 N2=N2+(M+1)                                                            01720000
C                                                                            01730000
CC    PRINT EVALUATION OF CALSSIFICATION FUNCTIONS FOR EACH OBSERVATION      01740000
C                                                                            01750000
      WRITE (6,12)                                                           01760000
      N1=1                                                                   01770000
      N2=N(1)                                                                01780000
      DO 210 I=1,K                                                           01790000
      WRITE (6,13) I                                                         01800000
      L=0                                                                    01810000
      DO 190 J=N1,N2                                                         01820000
      L=L+1                                                                  01830000
  190 WRITE (6,14) L,P(J),LG(J)                                              01840000
      IF(I-K) 200, 100, 100                                                  01850000
  200 N1=N1+N(I)                                                             01860000
      N2=N2+N(I+1)                                                           01870000
  210 CONTINUE                                                               01880000
      END                                                                    01890000
```

Programmet MDISC kalder 3 rutiner fra SSP-pakken, nemlig DMATX, MINV og DISCR. Disse er p.t. alle lagt ind under WATFIV compileren, hvilket muliggør en meget hurtig afvikling af et program.

En udskrift er vist p. 7.40 og 7.41.

I den anførte version udføres en diskriminantanalyse for indtil 5 populationer, hvor dimensionen af observationsvektorerne er højst 10. Der må højst være 250 observationer i alt. Hvis ens problem ikke kan honorere disse krav, må man ændre i programmet, som det er anført i Comment Statements eller p. 427 i SSP manualen.

Der kræves blot et enkelt styrekort, der udfyldes som følger

Kolonne	Indhold
1 - 6	Navn på analyse (gerne alfamerisk)
7 - 8	Antal populationer
9 - 10	Dimension af enkeltobservation
11 - 15	Antal observationer fra 1.ste pop.
16 - 20	" " " 2.den "
21 - 25	" " " 3.die "
.	.
.	.
.	.

Hvis vi benævner observationerne

$$
\begin{array}{ll}
\text{Pop. 1} & \left\{ \begin{array}{ll} X_{111}, \cdots, X_{11p} & \leftarrow \text{obs. 1 i pop. 1} \\ \vdots \quad\quad \vdots & \\ X_{1n_1 1}, \cdots, X_{1n_1 p} & \leftarrow \text{obs. } n_1 \text{ i pop. 1} \end{array} \right. \\
\vdots & \\
\text{Pop. k} & \left\{ \begin{array}{ll} X_{k11}, \cdots, X_{k1p} & \leftarrow \text{obs. 1 i pop. k} \\ \vdots \quad\quad \vdots & \\ X_{kn_k 1} \cdots X_{kn_k p} & \leftarrow \text{obs. } n_k \text{ i pop. k} \end{array} \right.
\end{array}
$$

skal de indlæses rækkevis. Hvis en enkelt række ikke kan stå på et enkelt hulkort, fortsættes på et nyt. Hver række skal dog starte på et nyt kort.

Data indlæses efter format statement 5, d.v.s.: i den anførte udskrift

```
5 FORMAT (F 4.0, 4x, F 4.0, 4x) .
```

Vi vil nu vise et par eksempler på kørsel med MDISC.

Eksempel 7.9 Vi betragter data fra den p. 7.20 omtalte undersøgelse af Fisher. Det drejer sig om målinger af bægerbladets længde og bredde og kronbladets længde og bredde på 3 Irisarter, Iris Setosa, Iris Versicolor og Iris Virginica. Der er foretaget 50 målinger fra hver population.

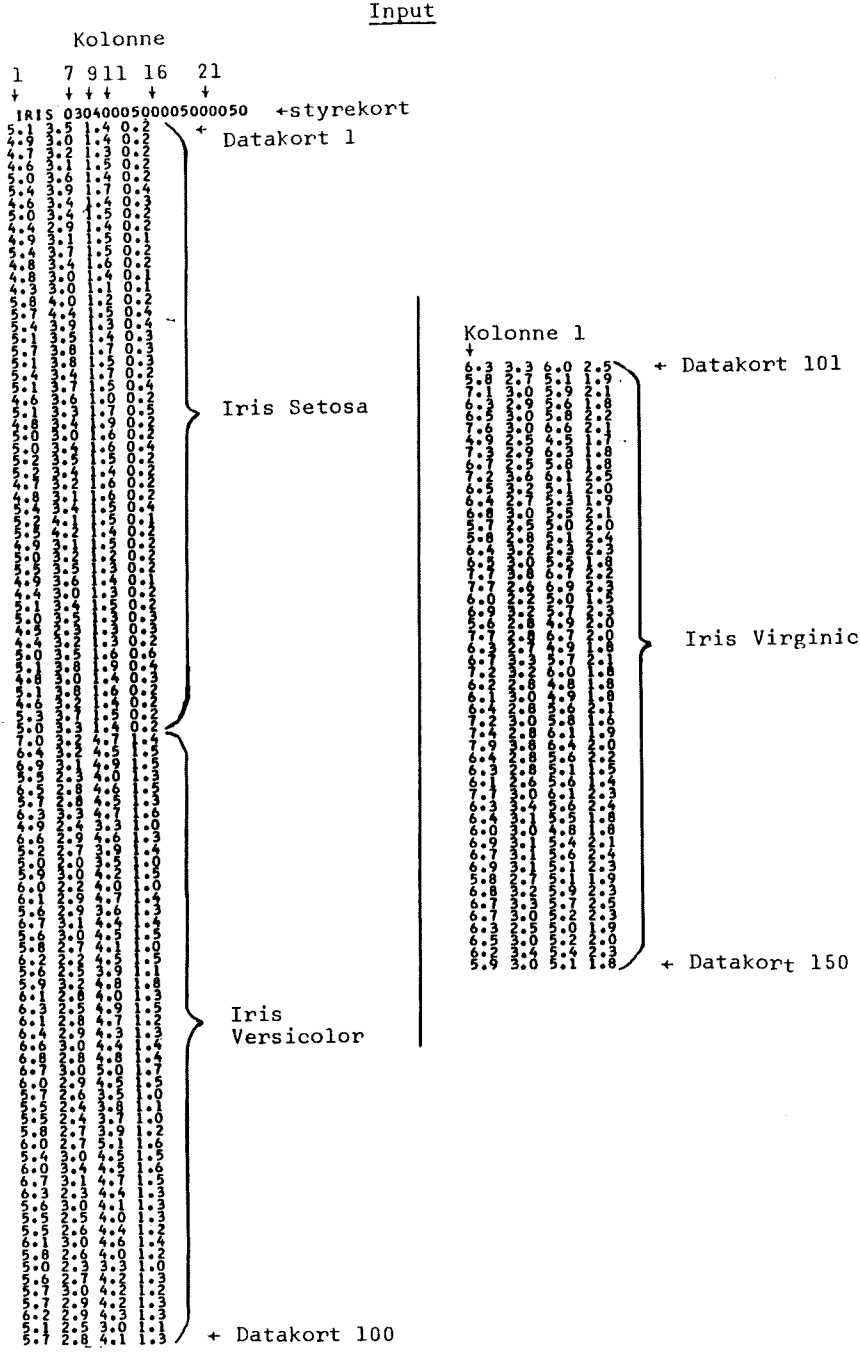
Format statement nr. 5 ændres til

```
FORMAT(4F4.0) .
```

Input er anført på p. 7.44. Kørslen er benævnt "Iris". De øvrige tal på styrekortet er åbenbare.

Output er anført på p. 7.45-47.

Input


Kolonne
1 7 9 11 16 21
IRIS 03040005000050000 ←styrekort
← Datakort 1
Iris Setosa
Iris Versicolor
← Datakort 100
Kolonne 1
← Datakort 101
Iris Virginica
← Datakort 150

Output

```
DISCRIMINANT ANALYSIS..... IRIS

  NUMBER OF GROUPS            3
  NUMBER OF VARIABLES         4
  SAMPLE SIZES...
        GROUP
          1          50
          2          50
          3          50


GROUP  1  MEANS
      5.00599        3.42799        1.46200        0.24600   ← x̄1'

GROUP  2  MEANS
      5.93598        2.76999        4.25999        1.32599   ← x̄2'

GROUP  3  MEANS
      6.58797        2.97399        5.55199        2.02599   ← x̄3'

POOLED DISPERSION MATRIX

ROW  1
      0.26500        0.09272        0.16751        0.03840
ROW  2
      0.09272        0.11539        0.05524        0.03271
ROW  3
      0.16751        0.05524        0.18519        0.04266
ROW  4
      0.03840        0.03271        0.04266        0.04188

                                                            } Σ̂

COMMON MEANS
      5.84331        3.05733        3.75799        1.19933   ← x̄'

GENERALIZED MAHALANOBIS D-SQUARE     4774.18300


DISCRIMINANT FUNCTION  1
     CONSTANT     *     COEFFICIENTS
    -85.21048     *          23.54451       23.58791      -16.43088      -17.39848
DISCRIMINANT FUNCTION  2
     CONSTANT     *     COEFFICIENTS
    -71.75427     *          15.69841        7.07245        5.21155        6.43410
DISCRIMINANT FUNCTION  3
     CONSTANT     *     COEFFICIENTS
   -103.27000     *          12.44600        3.68520       12.76666       21.07910
```

EVALUATION OF CLASSIFICATION FUNCTIONS FOR EACH OBSERVATION

GROUP 1

OBSERVATION	PROBABILITY ASSOCIATED WITH LARGEST DISCRIMINANT FUNCTION	LARGEST FUNCTION NO.
1	1.00000	1
2	1.00000	1
3	1.00000	1
4	1.00000	1
5	1.00000	1
6	1.00000	1
7	1.00000	1
8	1.00000	1
9	1.00000	1
10	1.00000	1
11	1.00000	1
12	1.00000	1
13	1.00000	1
14	1.00000	1
15	1.00000	1
16	1.00000	1
17	1.00000	1
18	1.00000	1
19	1.00000	1
20	1.00000	1
21	1.00000	1
22	1.00000	1
23	1.00000	1
24	1.00000	1
25	1.00000	1
26	1.00000	1
27	1.00000	1
28	1.00000	1
29	1.00000	1
30	1.00000	1
31	1.00000	1
32	1.00000	1
33	1.00000	1
34	1.00000	1
35	1.00000	1
36	1.00000	1
37	1.00000	1
38	1.00000	1
39	1.00000	1
40	1.00000	1
41	1.00000	1
42	1.00000	1
43	1.00000	1
44	1.00000	1
45	1.00000	1
46	1.00000	1
47	1.00000	1
48	1.00000	1
49	1.00000	1
50	1.00000	1

GROUP 2

OBSERVATION	PROBABILITY ASSOCIATED WITH LARGEST DISCRIMINANT FUNCTION	LARGEST FUNCTION NO.
1	0.99989	2
2	0.99926	2
3	0.99581	2
4	0.99964	2
5	0.99559	2
6	0.99850	2
7	0.98584	2
8	1.00000	2
9	0.99988	2
10	0.99950	2
11	1.00000	2
12	0.99923	2
13	1.00000	2
14	0.99433	2
15	1.00000	2
16	0.99996	2
17	0.98065	2
18	1.00000	2
19	0.95958	2
20	1.00000	2
21	0.74674	2
22	0.99999	2
23	0.81596	2
24	0.99957	2
25	0.99998	2
26	0.99992	2
27	0.99825	2
28	0.68926	2
29	0.99252	2
30	1.00000	2
31	1.00000	2
32	1.00000	2
33	1.00000	2
34	0.85659	2
35	0.96356	2
36	0.99404	2
37	0.99822	2
38	0.99946	2
39	0.99995	2
40	0.99982	2
41	0.99939	2
42	0.99809	2
43	0.99999	2
44	1.00000	2
45	0.99970	2
46	0.99998	2
47	0.99989	2
48	0.99995	2
49	1.00000	2
50	0.99993	2

GROUP 3

OBSERVATION	PROBABILITY ASSOCIATED WITH LARGEST DISCRIMINANT FUNCTION	LARGEST FUNCTION NO.
1	1.00000	3
2	0.99892	3
3	0.99997	3
4	0.99893	3
5	1.00000	3
6	1.00000	3
7	0.95137	3
8	0.99986	3
9	0.99978	3
10	1.00000	3
11	0.98695	3
12	0.99833	3
13	0.99980	3
14	0.99981	3
15	1.00000	3
16	0.99997	3
17	0.99392	3
18	1.00000	3
19	1.00000	3
20	0.77918	3
21	0.99992	3
22	0.99917	3
23	1.00000	3
24	0.90287	3
25	0.99991	3
26	0.99732	3
27	0.81160	3
28	0.86574	3
29	0.99999	3
30	0.89630	3
31	0.99986	3
32	0.99948	3
33	1.00000	3
34	0.72943	2
35	0.93397	3
36	1.00000	3
37	1.00000	3
38	0.99383	3
39	0.60745	3
40	0.99917	3
41	1.00000	3
42	0.99957	3
43	0.99892	3
44	1.00000	3
45	1.00000	3
46	0.99993	3
47	0.99410	3
48	0.99685	3
49	0.99999	3
50	0.98246	3

← Denne observation fra population 3 ligger nærmere $\hat{\mu}_2$ end $\hat{\mu}_3$

Der er mindre "tvivl" om "klassifikationen" for observation 45 end for observation 50

Vi ser e.g., at den generaliserede Mahalanobis D^2 er

$$4774.1830 \ .$$

Den skal sammenlignes med en fraktil i en $\chi^2(4(3-1)) = \chi^2(8)$-fordeling, og vi ser, at vi forkaster hypotesen $\underline{\mu}_1 = \underline{\mu}_2 = \underline{\mu}_3$ på alle niveauer $\alpha > 0.005$. Vi vil derfor gå ud fra, at det er rimeligt at bruge målinger på bægerblade og kronblade til at skelne mellem de 3 Irisarter. Diskriminantfunktionerne fremgår af output p. 7.45.

□

Eksempel 7.10 Vi betragter nu problemet, der er anført i eksemplet p. 7.23, hvor vi vil skelne mellem Versicolor og Setosa. Hvilke ændringer skal vi foretage i vort program og vore data? Vi fjerner selvfølgelig Virginica data fra datakortene. Dernæst ændres statement 5 tilbage til det oprindelige:

```
1 FORMAT(A4,A2,2I2,12I5/(14I5))                                    00850000
2 FORMAT(27H1DISCRIMINANT ANALYSIS.....,A4,A2/19HO  NUMBER OF GROUPS00860000
 1,7X,I3/22H   NUMBER OF VARIABLES,I7/17H   SAMPLE SIZES../12X,5HGRO00870000
 2UP)                                                              00880000
3 FORMAT(12X,I3,8X,I4)                                             00890000
4 FORMAT(1HO)                                                      00900000
5 FORMAT(F4.0,4X,F4.0,4X)
6 FORMAT(6HOGROUP,I3,7H  MEANS/(8F15.5))                           00920000
7 FORMAT(1HO/25H POOLED DISPERSION MATRIX)                         00930000
8 FORMAT(4HOROW,I3/(8F15.5))                                       00940000
9 FORMAT(1HO//13H COMMON MEANS/(8F15.5))                           00950000
```

Dette FORMAT statement springer tal nr. 2 og tal nr. 4 over på hvert hulkort, d.v.s.: vi indlæser kun længdemålinger. De nødvendige ændringer i styrekortet er

```
1    7 9  11   16   20  ← kolonne nr.
+    + +  +    +    +
 IRIS 02020005000050     ← styrekort
5.1  3.5  1.4  0.2
4.9  3.0  1.4  0.2
4.7  3.2  1.3  0.2
4.6  3.1  1.5  0.2
5.0  3.6  1.4  0.2
5.4  3.9  1.7  0.4
4.6  3.4  1.4  0.3
5.0  3.4  1.5  0.2
4.4  2.9  1.4  0.2
4.9  3.1  1.5  0.1
      .
      .
      .
```

Vi får da et output som vist nedenfor.

Vi ser, at den generaliserede Mahalanobis afstand er 1918.33500. Den ordinære Mahalanobis afstand mellem de to populationer er derfor

$$D = \frac{50+50}{50 \cdot 50} \cdot 1918.33500$$

$$= 76.7334 \ ,$$

og denne størrelse kan så anvendes i de videre beregninger p. 7.23. (Den der anførte afstand er 76.7082. Afvigelsen skyldes afrundingsfejl).

```
DISCRIMINANT ANALYSIS..... IRIS
  NUMBER OF GROUPS              2
  NUMBER OF VARIABLES           2
  SAMPLE SIZES..
        GROUP
          1                    50
          2                    50


GROUP  1  MEANS
       5.00599          1.46200

GROUP  2  MEANS
       5.93598          4.25999

POOLED DISPERSION MATRIX

ROW  1
       0.19534          0.09963

ROW  2
       0.09963          0.12549


COMMON MEANS
       5.47098          2.86099

GENERALIZED MAHALANOBIS D-SQUARE      1918.33500


DISCRIMINANT FUNCTION  1
     CONSTANT    *     COEFFICIENTS
    -72.11633    *          33.07930      -14.61162
DISCRIMINANT FUNCTION  2
     CONSTANT    *     COEFFICIENTS
   -100.36300    *          21.97012       16.50519
```

EVALUATION OF CLASSIFICATION FUNCTIONS FOR EACH OBSERVATION

GROUP 1

OBSERVATION	PROBABILITY ASSOCIATED WITH LARGEST DISCRIMINANT FUNCTION	LARGEST FUNCTION NO.
1	1.00000	1
2	1.00000	1
3	1.00000	1
4	1.00000	1
5	1.00000	1
6	1.00000	1
7	1.00000	1
8	1.00000	1
9	1.00000	1
10	1.00000	1
11	1.00000	1
12	1.00000	1
13	1.00000	1
14	1.00000	1
15	1.00000	1
16	1.00000	1
17	1.00000	1
18	1.00000	1
19	1.00000	1
20	1.00000	1
21	1.00000	1
22	1.00000	1
23	1.00000	1
24	1.00000	1
25	1.00000	1
26	1.00000	1
27	1.00000	1
28	1.00000	1
29	1.00000	1
30	1.00000	1
31	1.00000	1
32	1.00000	1
33	1.00000	1
34	1.00000	1
35	1.00000	1
36	1.00000	1
37	1.00000	1
38	1.00000	1
39	1.00000	1
40	1.00000	1
41	1.00000	1
42	1.00000	1
43	1.00000	1
44	1.00000	1
45	1.00000	1
46	1.00000	1
47	1.00000	1
48	1.00000	1
49	1.00000	1
50	1.00000	1

GROUP 2

OBSERVATION	PROBABILITY ASSOCIATED WITH LARGEST DISCRIMINANT FUNCTION	LARGEST FUNCTION NO.
1	1.00000	2
2	1.00000	2
3	1.00000	2
4	1.00000	2
5	1.00000	2
6	1.00000	2
7	1.00000	2
8	1.00000	2
9	1.00000	2
10	1.00000	2
11	1.00000	2
12	1.00000	2
13	1.00000	2
14	1.00000	2
15	1.00000	2
16	1.00000	2
17	1.00000	2
18	1.00000	2
19	1.00000	2
20	1.00000	2
21	1.00000	2
22	1.00000	2
23	1.00000	2
24	1.00000	2
25	1.00000	2
26	1.00000	2
27	1.00000	2
28	1.00000	2
29	1.00000	2
30	1.00000	2
31	1.00000	2
32	1.00000	2
33	1.00000	2
34	1.00000	2
35	1.00000	2
36	1.00000	2
37	1.00000	2
38	1.00000	2
39	1.00000	2
40	1.00000	2
41	1.00000	2
42	1.00000	2
43	1.00000	2
44	1.00000	2
45	1.00000	2
46	1.00000	2
47	1.00000	2
48	1.00000	2
49	0.99979	2
50	1.00000	2

Referencer til kapitel 7

Anderson, T. W.: An Introduction to Multivariate Statistical Analysis. John Wiley & Sons, New York 1958.

Cooley, W. W. & Lohnes, P. R.: Multivariate Data Analysis. John Wiley & Sons, New York 1971.

Dixon, J. W. (ed.): Biomedical Computer Programs. University of California Press, Los Angeles 1973.

System/360 Scientific Subroutine Package. Version III (fifth ed.). International Business Machines Corporation 1970.

KAPITEL 8

Principale komponenter
kanoniske variable og korrelationer
og
faktoranalyse

I dette kapitel skal vi give en indledende oversigt over nogle af de metoder, der bruges til at blotlægge den underliggende struktur i et flerdimensionalt datamateriale.

De principale komponenter svarer blot til resultaterne af en egenværdianalyse af dispersionsmatricen for en flerdimensional, stokastisk variabel. Metoden daterer sig tilbage til tiden omkring århundredskiftet (Karl Pearson), men først i trediverne fik den sin præcise udformning af Harold Hotelling.

Faktoranalysen er oprindelig udviklet af psykologer - Spearman (1904) og Thurstone i de første decennier af indeværende sekel. Dette har bevirket, at terminologien i (beklageligt) høj grad er præget af psykologers terminologi. Omkring 1940 udviklede Lawley maximum likelihood løsninger til problemer i faktoranalysen - arbejder, der senere bl.a. er fulgt op af Jöreskog, og herved introduceredes faktoranalysen som en "statistisk metode".

De kanoniske variable og korrelationer daterer sig også tilbage til Harold Hotelling. Begreberne minder meget om principale komponenter, blot undersøger vi nu samvariationen mellem to variable i stedet for at transformere en enkelt.

I denne fremstilling vælger vi at anskue problemerne ud fra en dataanalytisk synsvinkel, i.e. vi bekymrer os mere om at bestemme de empiriske strukturer end om at give teoretiske begrundelser for, at de inferenser, vi ønsker at - eller rettere sagt - som vi vil drage, er rimelige.

8.1 Principale komponenter

8.1.1 Definition og simple egenskaber

Vi betragter en flerdimensional, stokastisk variabel

$$\underline{X} = \begin{pmatrix} X_1 \\ \cdot \\ \cdot \\ \cdot \\ X_k \end{pmatrix},$$

der har dispersionsmatricen

$$D(\underline{X}) = \underline{\underline{\Sigma}} ,$$

og som uden tab af generalitet kan antages at have middelværdien $\underline{0}$.

Vi ordner egenværdierne i $\underline{\underline{\Sigma}}$ i dalende rækkefølge og benævner dem

$$\lambda_1 \geq \cdots \geq \lambda_k .$$

De tilsvarende ortonormerede egenvektorer benævnes

$$\underline{p}_1, \cdots, \underline{p}_k ,$$

og vi definerer den ortogonale matrix $\underline{\underline{P}}$ ved

$$\underline{\underline{P}} = (\underline{p}_1 \cdots \underline{p}_k) .$$

Vi har da følgende

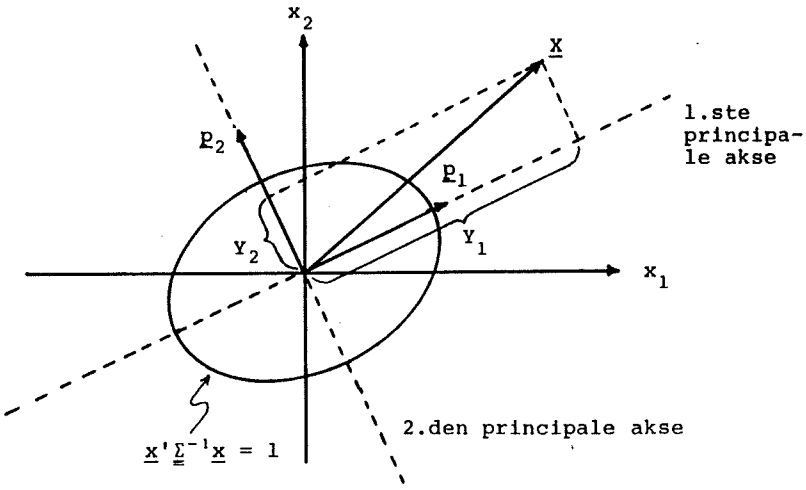
Definition 8.1 Ved den i'te principale akse for $\underline{X}$ forstås retningen hørende til egenvektoren $\underline{p}_i$ svarende til den i'te største egenværdi.

Definition 8.2 Ved den i'te principale komponent af $\underline{X}$ forstås $\underline{X}$'s projektion $Y_i = \underline{p}_i'\underline{X}$ på den i'te principale akse.

Vektoren

$$\underline{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_k \end{pmatrix} = \underline{P}'\underline{X}$$

kaldes vektoren af principale komponenter.



Situationen er anskueliggjort geometrisk i ovenstående figur, hvor vi har indtegnet den til kovariansstrukturen svarende enhedsellipsoide, i.e. ellipsoiden med ligningen

$$\underline{x}' \underline{\underline{\Sigma}}^{-1} \underline{x} = 1 \ .$$

Det ses da, at de principale akser netop bliver hovedakserne i denne ellipsoide.

Der gælder nu en række sætninger om egenskaberne ved de principale komponenter. De fleste af disse sætninger er statistiske reformuleringer af en række af de resultater angående symmetriske, positivt semidefinitte matricer, som er nævnt i kapitel 1.

Sætning 8.1 De principale komponenter er ukorrelerede, og variansen på den i'te komponent er λ_i, i.e. den i'te største egenværdi.

Bevis Vi har ifølge sætningerne 2.5 og 1.10

$$D(\underline{Y}) = D(\underline{\underline{P}}' \underline{X}) = \underline{\underline{P}}' \underline{\underline{\Sigma}} \underline{\underline{P}} = \underline{\underline{\Lambda}}$$

$$\begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \lambda_k \end{pmatrix} ,$$

og resultatet følger umiddelbart. □

Endvidere gælder

Sætning 8.2 Den generaliserede varians af de principale komponenter er lig den generaliserede varians af de oprindelige observationer.

Bevis Ifølge definitionen p. 2.50 haves

$$GV(\underline{X}) = \det \underline{\underline{\Sigma}}$$

og

$$GV(\underline{Y}) = \det \underline{\underline{\Lambda}} = \lambda_1 \cdots \lambda_k \ ,$$

og af relationen p. 1.48 følger umiddelbart, at $GV(\underline{X}) = GV(\underline{Y})$.

□

Et lignende resultat er

Sætning 8.3 Summen af de oprindelige variables varians er lig summen af de principale komponenters varians, i.e.

$$\sum_i V(X_i) = \sum_i V(Y_i)$$

Bevis Da

$$\Sigma V(X_i) = \text{tr } \underline{\underline{\Sigma}}$$

og

$$\Sigma V(Y_i) = \text{tr } \underline{\underline{\Lambda}}$$

følger resultatet af bemærkningen

□

Endelig har vi

Sætning 8.4 Den første principale komponent er den linearkombination (med normerede koefficienter) af de oprindelige variable, der har den største varians. Den m'te principale komponent er den linearkombination (med normerede koefficienter) af de oprindelige variable, som er ukorreleret med de m-1 første principale komponenter og har størst varians. Udtrykt formelt

$$\sup_{||b||=1} V(\underline{b}'\underline{X}) = \lambda_1 ,$$

og supremum antages for $\underline{b} = \underline{p}_1$. Endvidere er

$$\sup_{\underline{b} \perp \underline{p}_1, \cdots, \underline{p}_{m-1} \atop ||\underline{b}||=1} V(\underline{b}'\underline{X}) = \lambda_m ,$$

og supremum antages for $\underline{b} = \underline{p}_m$.

Bevis Da

$$V(\underline{b}'\underline{X}) = \underline{b}'\underline{\underline{\Sigma}}\,\underline{b} ,$$

og

$$\begin{aligned} Cov(Y_i, \underline{b}'\underline{X}) &= Cov(\underline{p}_i'\underline{X}, \underline{b}'\underline{X}) = \underline{p}_i'\underline{\underline{\Sigma}}\,\underline{b} \\ &= \lambda_i\, \underline{p}_i'\underline{b} , \end{aligned}$$

hvorfor

$$Cov(Y_i, \underline{b}'\underline{X}) = 0 \Leftrightarrow \underline{p}_i \perp \underline{b} ,$$

er sætningen blot en omformulering af sætning 1.15 p. 1.37.

□

Bemærkning Af sætningen får vi altså, at hvis vi søger den linearkombination af de oprindelige variable, der forklarer mest af variationen i disse, da er den første principale komponent løsningen. Søger vi de m variable, der forklarer mest muligt af den oprindelige variation, da er løsningen de m første principale komponenter. Et mål for, hvor godt disse beskriver den oprindelige variation, fås ved hjælp af sætningerne 1 og 3, der giver, at de m første principale komponenter beskriver brøkdelen

$$\frac{\lambda_1+\cdots+\lambda_m}{\lambda_1+\cdots+\lambda_m+\cdots+\lambda_k}$$

af den oprindelige variation.

Et mere kvalificeret mål for, hvor stor "genskabelsesevnen" er, fås ved at forsøge at rekonstruere det oprindelige $\underline{X}$ ud fra vektoren

$$\underline{Y}^* = (Y_1,\cdots,Y_m,0,\cdots,0)' \ .$$

Da

$$\underline{Y} = \underline{\underline{P}}'\underline{X} \leftrightarrow \underline{X} = \underline{\underline{P}}\ \underline{Y} \ ,$$

vil det være nærliggende at forsøge med

$$\underline{X}^* = \underline{\underline{P}}\ \underline{Y}^* \ .$$

Vi finder

$$D(\underline{X}^*) = \underline{\underline{P}}\ D(\underline{Y}^*)\underline{\underline{P}}'$$

$$= (\underline{p}_1\cdots\underline{p}_k)\begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_m \end{pmatrix}\begin{pmatrix} \underline{p}_1' \\ \vdots \\ \underline{p}_k' \end{pmatrix}$$

$$= \lambda_1\underline{p}_1\underline{p}_1' + \cdots + \lambda_m\underline{p}_m\underline{p}_m' \ .$$

Spektralfremstillingen af $\underline{\underline{\Sigma}}$ er (p. 1.39)

$$\underline{\underline{\Sigma}} = \lambda_1\underline{p}_1\underline{p}_1' + \cdots + \lambda_m\underline{p}_m\underline{p}_m' + \cdots + \lambda_k\underline{p}_k\underline{p}_k' \ ,$$

hvorfor

$$\underline{\underline{\Sigma}} - D(\underline{X}^*) = \lambda_{m+1}\underline{p}_{m+1}\underline{p}_{m+1}' + \cdots + \lambda_k\underline{p}_k\underline{p}_k' \ .$$

Hvis der er stor forskel på egenværdierne, vil de mindste være negligible, og forskellen mellem den oprindelige dispersionsmatrix og den "rekonstruerede" variabels dispersionsmatrix er derfor ringe.

8.1.2 Estimation og testning

Hvis dispersionsmatricen ikke er kendt, men estimeret på basis af n målinger, skønner man over de principale komponenter og

deres varianser ved blot at regne på den estimerede dispersionsmatrix, som om den var kendt. Hvis alle egenværdier i $\underline{\underline{\Sigma}}$ er forskellige, kan det vises, at de egenværdier og egenvektorer, vi får frem på denne måde, er maximum likelihood skøn over de sande parametre, se f. eks. T.W. Anderson (1958).

Der rejser sig dog et almindeligt problem her, idet det kan vises, at de principale komponenter ikke er uafhængige af de måleskalaer, vore oprindelige variable er målt i. Derfor vælger man tit kun at betragte normerede variable, i.e.

$$Y_{\ell i} = \frac{X_{\ell i} - \bar{X}_{\ell}}{\sqrt{\sum_i (\bar{X}_{\ell i} - \bar{X}_{\ell})^2/(n-1)}} ,$$

hvor

$$\underline{X}_i = \begin{pmatrix} X_{1i} \\ \vdots \\ X_{ki} \end{pmatrix} , \quad i = 1, \cdots, n .$$

Denne overgang svarer til, at vi analyserer den empiriske korrelationsmatrix i stedet for den empiriske dispersionsmatrix.

Hvis man beslutter sig til kun at medtage en del af de principale komponenter i den videre analyse, kan man f. eks. vælge en strategi, som at man tager så mange af komponenterne med, at de svarer til 90% af den totale variation.

Et andet kriterium vil være at teste hypoteser som

$$H_o: \lambda_1 \geq \cdots \geq \lambda_m \geq \lambda_{m+1} = \cdots = \lambda_k$$

mod alternativet, at der forekommer et skarpt ulighedstegn blandt de k-m sidste egenværdier.

Hvis vi arbejder med den estimerede dispersionsmatrix $\hat{\underline{\underline{\Sigma}}}$, bliver teststørrelsen

$$Z_1 = -n' \log_e \frac{\det \hat{\underline{\underline{\Sigma}}}}{\hat{\lambda}_1 \cdots \hat{\lambda}_m \cdot \hat{\lambda}^{k-m}} = -n' \log_e \frac{\hat{\lambda}_{m+1} \cdots \hat{\lambda}_k}{\hat{\lambda}^{k-m}} ,$$

hvor

$$n' = n - m - \frac{1}{6}\left(2(k-m)+1+\frac{2}{k-m}\right) ,$$

og

$$\hat{\lambda} = (\operatorname{tr} \hat{\underline{\underline{\Sigma}}} - \hat{\lambda}_1 - \cdots - \hat{\lambda}_m)/(k-m) = (\hat{\lambda}_{m+1} + \cdots + \hat{\lambda}_k)/(k-m) .$$

Det kritiske område ved test på niveau α er approximativt

$$\{(\underline{x}_1 \cdots, \underline{x}_n) \mid z_1 > \chi^2(\tfrac{1}{2}(k-m+2)(k-m-1))_{1-\alpha}\} .$$

Regner vi i stedet på den estimerede korrelationsmatrix $\hat{\underline{\underline{R}}}$, fås kriteriet

$$Z_2 = -n \log_e \frac{\det \hat{\underline{\underline{R}}}}{\hat{\lambda}_1 \cdots \hat{\lambda}_m \cdot \hat{\lambda}^{k-m}} = -n \log_e \frac{\hat{\lambda}_{m+1} \cdots \hat{\lambda}_k}{\hat{\lambda}^{k-m}} ,$$

hvor

$$\hat{\lambda} = (k - \hat{\lambda}_1 - \cdots - \hat{\lambda}_m)/(k-m) = (\hat{\lambda}_{m+1} + \cdots + \hat{\lambda}_k)/(k-m) .$$

Det kritiske område bliver ved test på niveau α approximativt lig

$$\{\underline{x}_1, \cdots, \underline{x}_n \mid z_2 > \chi^2(\tfrac{1}{2}(k-m+2)(k-m-1))_{1-\alpha}\} .$$

Det må dog her indskærpes, at denne approximation er langt dårligere end den tilsvarende for dispersionsmatricen.

En diskussion af ovennævnte tests kan findes i Lawley & Maxwell (1963).

Vi giver nu et eksempel.

Eksempel 8.1 Eksemplet er baseret på et eksempel fra Davis (1973) p. 486. Udgangsmaterialet er målinger af 7 variable på 25 kasser med tilfældigt genererede sider. De 7 variable er

X_1: længste side

X_2: mellemste side

X_3: korteste side

X_4: længste diagonal

X_5: radius i omskreven kugle/radius i indskreven kugle

X_6: (længste side + mellemste side)/korteste side

X_7: overfladeareal/volumen.

I nedenstående tabel er vist et udsnit af målingerne af de 7 variable.

Kasse	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	3.760	3.660	0.540	5.275	9.768	13.741	4.782
2	8.590	4.990	1.340	10.022	7.500	10.162	2.130
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
24	8.210	3.080	2.420	9.097	3.753	4.657	1.719
25	9.410	6.440	5.110	12.495	2.446	3.103	0.914

Vi stiller os bl.a. det spørgsmål: Hvilke forhold ved en kasse er afgørende for, hvorledes vi opfatter dens størrelse?

Med henblik på besvarelsen af dette spørgsmål foretager vi en principal komponent-analyse af ovenstående datamateriale. Ved en sådan analyse håber vi at få belyst, hvorvidt de ovennævnte 7 variable, der alle på en eller anden måde er forbundet med

"størrelse" og "form", varierer frit i det 7-dimensionale talrum, eller om de mere eller mindre udpræget er koncentreret i nogle underrum.

Vi angiver først den empiriske dispersionsmatrix for de variable. Den er

$$\hat{\underline{\underline{\Sigma}}} = \begin{bmatrix} 5.400 & 3.260 & 0.779 & 6.391 & 2.155 & 3.035 & -1.996 \\ 3.260 & 5.846 & 1.465 & 6.083 & 1.312 & 2.877 & -2.370 \\ 0.779 & 1.465 & 2.774 & 2.204 & -3.839 & -5.167 & -1.740 \\ 6.391 & 6.083 & 2.204 & 9.107 & 1.610 & 2.782 & -3.283 \\ 2.155 & 1.312 & -3.839 & 1.610 & 10.710 & 14.770 & 2.252 \\ 3.035 & 2.877 & -5.167 & 2.782 & 14.770 & 20.780 & 2.622 \\ -1.996 & -2.370 & -1.740 & -3.283 & 2.252 & 2.622 & 2.594 \end{bmatrix}$$

Dernæst bestemmes egenvektorerne og egenværdierne for $\hat{\underline{\underline{\Sigma}}}$. Egenværdierne er i dalende rækkefølge med samtidig angivelse af dels den brøkdel og dels den kumulerede brøkdel af den totale varians, egenværdierne bidrager med:

Egenværdi $\hat{\lambda}_i, i=1,\cdots,7$	Procentdel af total varians	Kumuleret procentdel af total varians
34.490	60.290	60.290
19.000	33.210	93.500
2.540	4.440	97.940
0.810	1.410	99.350
0.340	0.600	99.950
0.033	0.060	100.010
0.003	0.004	100.014

De afvigelser, som betinger, at den kumulerede sum overstiger 100%, er regneunøjagtighed ved bestemmelsen af egenværdierne.

De tilhørende egenvektorers koordinater er vist i omstående tabel.

Variabel	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$	$\hat{p}_4$	$\hat{p}_5$	$\hat{p}_6$	$\hat{p}_7$
X_1	0.164	0.422	0.645	-0.090	0.225	0.415	-0.385
X_2	0.142	0.447	-0.713	-0.050	0.395	0.066	-0.329
X_3	-0.173	0.257	-0.130	0.629	-0.607	0.280	-0.211
X_4	0.170	0.650	0.146	0.212	0.033	-0.403	0.565
X_5	0.546	-0.135	0.105	0.165	-0.161	-0.596	-0.513
X_6	0.768	-0.133	-0.149	-0.062	-0.207	0.465	0.327
X_7	0.073	-0.313	0.065	0.719	0.596	0.107	0.092

Det fremgår, at den første egenvektor, hvis retning tager højde for mere end 60% af den totale variation, især har numerisk store 5'te og 6'te koordinater. Dette bevirker, at den første principale komponent

$$Y_1 = 0.164\ X_1 + \cdots + 0.546\ X_5 + 0.768\ X_6 + 0.073\ X_7$$

er særlig følsom over for variationer i X_5 og X_6. Disse to variable, nemlig kvotienten mellem radius i den omskrevne radius i den indskrevne kugle samt kvotienten mellem summen af de to længste sider og den mindste side, har begge noget at gøre med, hvor "flad" en kasse er. Jo større disse to variable er, jo "fladere" er kassen. Den første principale komponent måler altså forskelle i "tykkelsen" af kasserne.

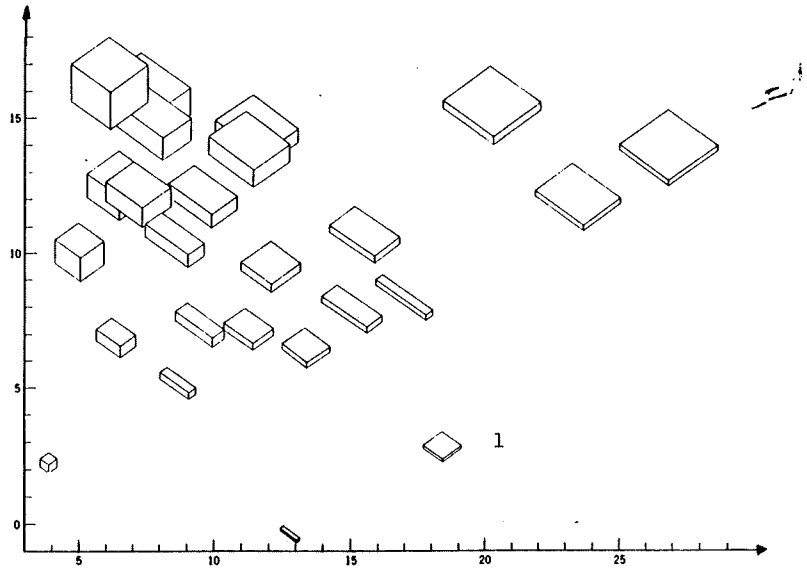
Den anden egenvektor har store positive koordinater på de 4 første pladser og ret stor negativ koordinat på sidste plads. Hvis den anden principale komponent

$$Y_2 = 0.422\,X_1 + 0.447\,X_2 + 0.257\,X_3 + 0.650\,X_4 + \cdots - 0.313\,X_7\,,$$

er meget stor, må en eller flere af $X_1, \cdots, X_4$ være store og X_7 lille. Nu gælder det, at terningen er den kasse, der for et givet volumen har den mindste overflade. Hvis en kasse derfor afviger meget fra en terning, vil den have stor X_7-værdi, og

dette vil medvirke til en kraftig reduktion af Y_2. En stor Y_2-værdi tyder derfor på, at de fleste sider er store og nogenlunde lige store. Vi konkluderer derfor, at Y_2 måler et mere generelt størrelsesbegreb.

I nedenstående figur har man afbildet kasserne i et koordinatsystem, hvis akser er de to første principale akser. Koordinaterne for den enkelte kasse bliver derfor værdien af den første og den anden principale komponent for den specifikke kasse.



For den første kasse finder vi e.g.

$$Y_1 = 0.164\ 3.760 + \cdots + 0.073\ 4.782 = 18.18$$
$$Y_2 = 0.422\ 3.760 + \cdots - 0.313\ 4.782 = 2.15\ .$$

I punktet med koordinater (18.18, 2.15) er der dernæst tegnet et billede af kasse nr. 1, etc.

Af denne graf fremgår også tydeligt den tolkning, vi har givet af de principale komponenters betydning. Til venstre i billedet - svarende til små værdier af komponent nr. 1 - har vi de tykkeste kasser og til højre de fladeste. Øverst i billedet - svarende til store værdier af komponent nr. 2 - har vi de store kasser og nederst de små.

Der synes til gengæld ikke at være nogen præcis skelnen mellem stavformede kasser og mere flade kasser. Denne skelnen kommer først frem, når vi også involverer den tredie principale komponent. Den er

$$Y_3 = 0.645\ X_1 - 0.713\ X_2 + \cdots + 0.065\ X_7\ .$$

Denne komponent lægger stor positiv vægt på variabel 1 - længden af den største side - og stor negativ vægt på længden af den næststørste side. I en udpræget stavformet kasse vil $X_1 >> X_2$, og derfor vil Y_3 være relativt stor for en sådan. Hvis grundfladen udspændt af de to største sider nærmer sig et kvadrat, vil Y_3 praktisk taget være 0 for den pågældende kasse.

De tre første principale komponenter tager altså højde for ca. 98% af den totale variation, og ved hjælp af disse er vi i stand til at splitte en kasses "størrelsesmæssige karakteristika" op i tre ukorrelerede komponenter: En, der angiver kassens fladhed (Y_1), en, der angiver et mere alment størrelsesbegreb (Y_2) og en, der angiver "graden af stavformethed" (Y_3). Hermed skulle det indledende spørgsmål: Hvad er "størrelsen af en kasse" være i det mindste delvist belyst.

□

Det næste eksempel er baseret på nogle undersøgelser af Agterberg et al. (se Agterberg (1973) p. 128).

Eksempel 8.2 Mount Albert peridotit intrusionen er en del af det Appalachiske ultramafiske bælte i Quebec provinsen. For en række indsamlede mineralstykker har man bestemt værdierne af følgende 4 variable:

X_1: mol% forsterit (= Mg-olivin)

X_2: mol% enstatit (= Mg-ortopyroxen)

X_3: enhedscelle dimensionen af chrom-spinel

X_4: specifikke vægtfylde af mineralstykket.

På basis af mellem 99 og 156 målinger har man dernæst estimeret følgende korrelationsmatrix mellem de variable:

$$\hat{\underline{\underline{R}}} = \begin{bmatrix} 1.00 & 0.32 & 0.41 & -0.31 \\ 0.32 & 1.00 & 0.68 & -0.38 \\ 0.41 & 0.68 & 1.00 & -0.36 \\ -0.31 & -0.38 & -0.36 & 1.00 \end{bmatrix}.$$

Det er her helt klart, at vi må analysere korrelationsmatricen og ikke dispersionsmatricen. Der er jo her tale om variable, der måles i vidt forskellige enheder, hvorfor vi må standardisere tallene.

Egenværdierne og de tilhørende egenvektorer er

$$\hat{\lambda}_1 = 2.25; \quad \hat{\underline{p}}_1 = \begin{bmatrix} 0.43 \\ 0.55 \\ 0.57 \\ -0.44 \end{bmatrix}$$

$$\hat{\lambda}_2 = 0.74; \quad \hat{\underline{p}}_2 = \begin{bmatrix} -0.66 \\ 0.49 \\ 0.37 \\ 0.44 \end{bmatrix}$$

$$\hat{\lambda}_3 = 0.70; \quad \hat{\underline{p}}_3 \quad \begin{bmatrix} 0.60 \\ -0.02 \\ 0.16 \\ 0.78 \end{bmatrix}$$

$$\hat{\lambda}_4 = 0.31; \quad \hat{\underline{p}}_4 = \begin{bmatrix} -0.14 \\ -0.68 \\ 0.72 \\ -0.06 \end{bmatrix} .$$

Alle egenvektorerne har rimeligt store koordinater på de fleste pladser, således at der ikke synes at være nogen mulighed for at give en intuitiv tolkning af de principale komponenter.

Den første principale komponent tager højde for 2.25/4 = 56.25% af den totale variation.

Det vil være interessant at få afgjort, hvorvidt de tre mindste egenvektorer for korrelationsmatricen kan antages at være af samme størrelsesorden.

Som teststørrelse anvendes

$$Z = -n \log \frac{0.74 \cdot 0.70 \cdot 0.31}{[(0.74+0.70+0.31)/3]^3} = 0.2120\ n ,$$

hvor n er antallet af observationer, korrelationsmatricen er baseret på. Dette antal er som nævnt ikke det samme for de forskellige korrelationskoefficienter, så derfor falder den teoretiske begrundelse for at anvende testet sådan set lidt væk. Hvis vi imidlertid ser bort fra disse problemer, bliver antallet af frihedsgrader i den χ^2-fordeling, vi skal sammenligne teststørrelsen med,

$$f = \frac{1}{2}(4-1+2)(4-1-1) = 5 .$$

Da

$$\chi^2(5)_{0.995} = 16.7 ,$$

og da 0.21 n for n af størrelsesordenen 100 er væsentligt større end denne værdi, vil det næppe være urimeligt at antage, at de tre mindste egenvektorer i (den "sande") korrelationsmatrix ikke er af samme størrelsesorden.

□

8.2 Kanoniske variable og kanoniske korrelationer

Vi skal i det følgende diskutere afhængighed mellem grupper af variable, hvor vi i det foregående afsnit alene så på afhængigheden (korrelationsstrukturen) mellem enkeltvariable.

Vi betragter en stokastisk variabel $\underline{X}$

$$\underline{X} \in N_{p+q}(\underline{\mu}, \underline{\underline{\Sigma}}) ,$$

hvor $p \leq q$, og hvor $\underline{X}$ og parametrene er spaltet som følger:

$$\underline{X} = \begin{pmatrix} \underline{X}_1 \\ \underline{X}_2 \end{pmatrix} , \quad \underline{\mu} = \begin{pmatrix} \underline{\mu}_1 \\ \underline{\mu}_2 \end{pmatrix} , \quad \underline{\underline{\Sigma}} = \begin{pmatrix} \underline{\underline{\Sigma}}_{11} & \underline{\underline{\Sigma}}_{12} \\ \underline{\underline{\Sigma}}_{21} & \underline{\underline{\Sigma}}_{22} \end{pmatrix} .$$

Hvis vi på basis af n målinger af $\underline{X}$ ønsker at undersøge, om $\underline{X}_1$ og $\underline{X}_2$ er uafhængige, kan dette som anført i kapitel 6 gøres ved at undersøge

$$\frac{\det(\underline{\underline{S}})}{\det(\underline{\underline{S}}_{11}) \det(\underline{\underline{S}}_{22})} ,$$

der er $U_{p,q,n-1-g}$ fordelt under H_o. Vi vil nu prøve at anskue problemet fra en lidt anden synsvinkel. Vi betragter to endimensionale variable Y_1 og Y_2 givet ved

$$U = \underline{a}'\underline{X}_1 \quad \text{og} \quad V = \underline{b}'\underline{X}_2 .$$

Da er

$$D\begin{pmatrix} U \\ V \end{pmatrix} = \begin{pmatrix} \underline{a}' \\ \underline{b}' \end{pmatrix} \underline{\underline{\Sigma}} \, (\underline{a}, \underline{b}) = \begin{bmatrix} \underline{a}'\underline{\underline{\Sigma}}_{11}\underline{a} & \underline{a}'\underline{\underline{\Sigma}}_{12}\underline{b} \\ \underline{b}'\underline{\underline{\Sigma}}_{21}\underline{a} & \underline{b}'\underline{\underline{\Sigma}}_{22}\underline{b} \end{bmatrix} ,$$

og korrelationen mellem U og V er

$$\rho(\underline{a},\underline{b}) = \frac{\underline{a}'\underline{\underline{\Sigma}}_{12}\underline{b}}{\sqrt{\underline{a}'\underline{\underline{\Sigma}}_{11}\underline{a}\ \underline{b}'\underline{\underline{\Sigma}}_{22}\underline{b}}} .$$

Nu er åbenbart

$$\underline{\underline{\Sigma}}_{12} = 0 \leftrightarrow \forall \underline{a},\underline{b} : \rho(\underline{a},\underline{b}) = 0 .$$

Acceptområdet for hypotesen $\rho(\underline{a},\underline{b}) = 0$ er af formen (jvf. kapitel 2)

$$r^2(\underline{a},\underline{b}) \leq r_\beta^2 ,$$

hvor $r(\underline{a},\underline{b})$ er den empiriske korrelationskoefficient, og r_β^2 er en passende fraktil i nulhypotesefordelingen. Vi får derfor accepteret $\underline{\underline{\Sigma}}_{12} = 0$, såfremt

$$\forall \underline{a},\underline{b} : r^2(\underline{a},\underline{b}) \leq r_\beta^2 ,$$

hvilket helt klart er ensbetydende med, at

$$\max_{\underline{a},\underline{b}} r^2(\underline{a},\underline{b}) \leq r_\beta^2 .$$

Vi er således nået frem til, at de to grupper er uafhængige, hvis den maksimale (empiriske) korrelationskoefficient mellem en linearkombination fra den første gruppe og en linearkombination fra den anden gruppe er tilpas lille. Denne maksimale korrelationskoefficient kaldes den første (empiriske) kanoniske korrelationskoefficient og de tilsvarende variable de første (empiriske) kanoniske variable.

Nu er det klart, at man ligesom i tilfældet med de principale komponenter kan "fortsætte" definitionen. Vi kan definere den anden kanoniske korrelationskoefficient som den maksimale korrelation mellem linearkombinationer af $\underline{X}_1$'erne og $\underline{X}_2$'erne, således at disse kombinationer er uafhængige af de foregående, etc. Helt stringent har vi

Definition 8.3 Lad $\underline{X} = \begin{pmatrix} \underline{X}_1 \\ \underline{X}_2 \end{pmatrix}$ være en stokastisk variabel, hvor $\underline{X}_1$ har p komponenter og $\underline{X}_2$ q komponenter ($p \leq q$). Det r'te par kanoniske variable er det par af linearkombinationer $U_r = \underline{\alpha}_r'\underline{X}_1$ og $V_r = \underline{\beta}_r'\underline{X}_2$, som hver har variansen 1, som er ukorrelerede med de foregående r-1 par af kanoniske variable, og som har maksimal korrelation. Korrelationen er den r'te kanoniske korrelation.

Tilbage står nu problemet med at bestemme de kanoniske variable og korrelationer.

Der gælder følgende sætning:

Sætning 8.5 Lad situationen være som i ovenstående definition, og lad $D(\underline{X}) = \underline{\underline{\Sigma}}$ være spaltet analogt

$$\underline{\underline{\Sigma}} = \begin{pmatrix} \underline{\underline{\Sigma}}_{11} & \underline{\underline{\Sigma}}_{12} \\ \underline{\underline{\Sigma}}_{21} & \underline{\underline{\Sigma}}_{22} \end{pmatrix}.$$

Da er den r'te kanoniske korrelation lig den r'te største rod λ_r af

$$\det \begin{pmatrix} -\lambda\underline{\underline{\Sigma}}_{11} & \underline{\underline{\Sigma}}_{12} \\ \underline{\underline{\Sigma}}_{21} & -\lambda\underline{\underline{\Sigma}}_{22} \end{pmatrix} = 0 ,$$

og koefficienterne i det r'te par kanoniske variable tilfredsstiller

$$\text{(i)} \quad \begin{pmatrix} -\lambda_r\underline{\underline{\Sigma}}_{11} & \underline{\underline{\Sigma}}_{12} \\ \underline{\underline{\Sigma}}_{21} & -\lambda_r\underline{\underline{\Sigma}}_{22} \end{pmatrix} \begin{pmatrix} \underline{\alpha}_r \\ \underline{\beta}_r \end{pmatrix} = \underline{0}$$

$$\text{(ii)} \quad \underline{\alpha}_r'\underline{\underline{\Sigma}}_{11}\underline{\alpha}_r = 1$$

$$\text{(iii)} \quad \underline{\beta}_r'\underline{\underline{\Sigma}}_{22}\underline{\beta}_r = 1 .$$

Bevis Der er tale om et maksimaliseringsproblem under bibetingelser, og man kan komme igennem ved at anvende en Lagrange-

multiplikator-teknik, se f. eks. Anderson (1958) p. 289.

Man kan også bestemme korrelationerne og koefficienterne ved at løse et egenværdiproblem, idet vi har

Sætning 8.6 Lad situationen være som i foregående sætning. Da gælder

$$(\underline{\underline{\Sigma}}_{12}\underline{\underline{\Sigma}}_{22}^{-1}\underline{\underline{\Sigma}}_{21} - \lambda_r^2\underline{\underline{\Sigma}}_{11})\underline{\alpha}_r = \underline{0}$$

$$\det(\underline{\underline{\Sigma}}_{12}\underline{\underline{\Sigma}}_{22}^{-1}\underline{\underline{\Sigma}}_{21} - \lambda_r^2\underline{\underline{\Sigma}}_{11}) = 0$$

respektive

$$(\underline{\underline{\Sigma}}_{21}\underline{\underline{\Sigma}}_{11}^{-1}\underline{\underline{\Sigma}}_{12} - \lambda_r^2\underline{\underline{\Sigma}}_{22})\underline{\beta}_r = \underline{0}$$

$$\det(\underline{\underline{\Sigma}}_{21}\underline{\underline{\Sigma}}_{11}^{-1}\underline{\underline{\Sigma}}_{12} - \lambda_r^2\underline{\underline{\Sigma}}_{22}) = 0$$

Bevis Forbigås, se f. eks. Anderson (1958).

Vedrørende estimationen er der intet særligt at tilføje. Indsættes maximum likelihood-skøn for $\underline{\underline{\Sigma}}$ i de foregående sætninger, fås maximum likelihood-skøn for parametrene. Hyppigst vil man nok indsætte det centrale skøn $\underline{\underline{S}}$, og man får da det, man kan kalde de empiriske værdier (engelsk: sample values) for de involverede parametre.

Der findes i de fleste systemer af standardprogrammer også programmer til evaluering af kanoniske korrelationer og - variable. Vi kan eksempelvis nævne BMDP6M: Canonical Correlation Analysis fra BMDP-pakken.

8.3 Faktoranalyse

Vi vender os nu igen mod analysen af korrelationsstrukturen for en enkelt flerdimensional variabel, men i modsætning til, hvad

der var tilfældet under afsnittet om principale komponenter, går vi her ud fra en underliggende model af strukturen.

8.3.1 Model og forudsætninger

Det forudsættes, at der foreligger en observation

$$\underline{X} = \begin{bmatrix} X_1 \\ \cdot \\ \cdot \\ \cdot \\ X_k \end{bmatrix},$$

der - hvis man vil bevare det historiske udviklingsforløb i tankerne - kan opfattes som en enkelt persons karakterer ved f. eks. k forskellige typer intelligenstest, eller, om man vil, en persons reaktion på k forskellige stimuli.

Man har så en model for, hvorledes man tænker sig, at disse reaktioner (karakterer) afhænger af nogle underliggende faktorer, eller mere specifikt, at

$$\underline{X} = \underline{\underline{A}}\,\underline{F} + \underline{G},$$

d.v.s. skrevet ud

$$\begin{bmatrix} X_1 \\ \vdots \\ X_k \end{bmatrix} = \begin{bmatrix} a_{11} \cdots a_{1m} \\ \vdots \quad\quad \vdots \\ a_{k1} \cdots a_{km} \end{bmatrix} \cdot \begin{bmatrix} F_1 \\ \vdots \\ F_m \end{bmatrix} + \begin{bmatrix} G_1 \\ \vdots \\ G_k \end{bmatrix}.$$

Her benævnes $\underline{F}$ vektoren af fælles faktorer (common factors), eller de kaldes faktorværdierne (factor scores). Disse er ikke observerbare. Eksempler på sådanne er egenskaber som rumlig intelligens, verbal intelligens etc.

$\underline{\underline{A}}$-matricens elementer kaldes faktorvægte (factor loadings), og de angiver de vægte, hvormed de enkelte faktorer indgår i beskrivelsen af de forskellige variable. Hvis man e.g. antager, at

F_l angiver rumlig intelligens og F_m verbal do., og at X_l er resultatet af en prøve af rumgeometrisk tilsnit og X_k resultatet af en læseprøve, ja da vil man selvsagt have, at a_{ll} er stor og a_{lm} lille og omvendt, at a_{kl} er lille og a_{km} stor, svarende til, at den rumlige intelligens er afgørende for personens karakter ved løsningen af rumlige opgaver, og analogt for den verbale intelligens.

Vektoren $\underline{G}$ kaldes vektoren af unike faktorer (unique factors), og den kan om ønsket tænkes sammensat af nogle specifikke faktorer (specific factors), faktorer som er specielle for netop disse konkrete tests, og så af "fejl", i.e. ikke-forklarede afvigelser. Disse faktorer er selvsagt heller ikke observerbare.

Det må her præciseres, at såvel $\underline{X}$ som $\underline{F}$ og $\underline{G}$ antages at være stokastiske. Der er derfor ikke tale om en generel lineær model med parametre $F_1, \cdots, F_m$.

For at gøre denne forskel helt tydelig vil vi derfor anføre modellen i det tilfælde, hvor vi har flere observationer $\underline{X}_1, \cdots, \underline{X}_n$. Vi har da de n modeller

$$\begin{bmatrix} X_{1i} \\ \vdots \\ X_{ki} \end{bmatrix} = \begin{bmatrix} a_{11} \cdots a_{1m} \\ \vdots \quad\quad \vdots \\ a_{k1} \cdots a_{km} \end{bmatrix} \begin{bmatrix} F_{1i} \\ \vdots \\ F_{mi} \end{bmatrix} + \begin{bmatrix} G_{1i} \\ \vdots \\ G_{ki} \end{bmatrix},$$

hvor vi bemærker, at det er $\underline{F}_i$ og $\underline{G}_i$, der ændres med observationerne $\underline{X}_i$.

Vi kan samle ovenstående modeller til

$$\begin{bmatrix} X_{11} \cdots X_{1n} \\ \vdots \quad\quad \vdots \\ X_{k1} \cdots X_{kn} \end{bmatrix} = \begin{bmatrix} a_{11} \cdots a_{1m} \\ \vdots \quad\quad \vdots \\ a_{k1} \cdots a_{km} \end{bmatrix} \begin{bmatrix} F_{11} \cdots F_{1n} \\ \vdots \quad\quad \vdots \\ F_{m1} \cdots F_{mn} \end{bmatrix} + \begin{bmatrix} G_{11} \cdots G_{1n} \\ \vdots \quad\quad \vdots \\ G_{k1} \cdots G_{kn} \end{bmatrix}.$$

Det forudsættes, at $\underline{F}$ og $\underline{G}$ er ukorrelerede, og at

$$D(\underline{F}) = \begin{pmatrix} 1 \cdots 0 \\ \vdots \quad \vdots \\ 0 \cdots 1 \end{pmatrix} = \underline{\underline{I}} = \underline{\underline{I}}_m \ ,$$

og

$$D(\underline{G}) = \begin{pmatrix} \delta_1 \cdots 0 \\ \vdots \quad \vdots \\ 0 \cdots \delta_k \end{pmatrix} = \underline{\underline{\Delta}} \ .$$

Endvidere forudsættes, at observationerne er standardiseret på en sådan måde, at $V(X_i) = 1$, $\forall i$, d.v.s. at dispersionsmatricen for $\underline{X}$ er lig dens korrelationsmatrix, som benævnes

$$D(\underline{X}) = \underline{\underline{R}} = \begin{pmatrix} 1 \quad \cdots r_{1k} \\ \vdots \quad \vdots \\ r_{k1} \cdots 1 \end{pmatrix} \ .$$

Af den oprindelige faktorligning fås ved hjælp af sætning 2.5 p. 2,5, at

$$\underline{\underline{R}} = \underline{\underline{A}} \, \underline{\underline{A}}' + \underline{\underline{\Delta}} \ .$$

Heraf udledes specielt, at vi for $j = 1, \cdots, k$ har

$$V(X_j) = a_{j1}^2 + \cdots + a_{jm}^2 + \delta_j = 1 \ .$$

Her indføres betegnelsen

$$h_j^2 = a_{j1}^2 + \cdots + a_{jm}^2 \ , \quad j = 1, \cdots, k \ .$$

Disse størrelser benævnes kommunaliteter (communalities), og h_j angiver, hvor stor en brøkdel af X_j's varians, der hidrører fra de m fælles faktorer.

Tilsvarende angiver δ_j den "uniqueness", der er i X_j's varians, i.e. den del af X_j's varians, der ikke hidrører fra de m fælles faktorer.

Endelig angiver den (i,j)'te faktorvægt korrelationen mellem den i'te variabel og den j'te faktor, d.v.s.

$$\mathrm{Cov}(X_i, F_j) = \mathrm{Cov}(\sum_{\nu} a_{i\nu}F_\nu + G_i, F_j) = a_{ij} \ .$$

Det kan vises (Dwyer (1939)), at

$$h_j^2 = a_{j1}^2 + \cdots + a_{jm}^2 \geq r_{j|1\cdots k}^2 \ ,$$

d.v.s. at den j'te kommunalitet altid er større end eller lig med kvadratet på den multiple korrelationskoefficient mellem X_j og de øvrige variable. Dette forekommer ikke urimeligt, når man erindrer, at denne størrelse netop angiver den brøkdel af X_j's varians, der forklares ved variationen i de øvrige X_i'er.

8.3.2 Estimation af faktorer (faktorvægte)

Vi går nu over til det mere konkrete problem at estimere faktorerne. Det, vi er interesseret i at bestemme, er $\underline{\underline{A}}$. Vi finder

$$\underline{\underline{A}}\,\underline{\underline{A}}' = \underline{\underline{R}} - \underline{\underline{\Delta}} \ .$$

Diagonalelementerne i denne matrix er

$$1 - \delta_j = h_j^2 \ , \quad j = 1, \cdots, k \ .$$

Disse kender vi ikke, men vi kan eventuelt estimere dem ved de multiple korrelationskoefficienters kvadrater. Indsættes disse, fås en matrix

$$\underline{\underline{V}} = \begin{bmatrix} r^2_{1|2\cdots k} & \cdots & r_{1k} \\ \vdots & & \vdots \\ r_{k1} & \cdots & r^2_{k|1\cdots k-1} \end{bmatrix} ,$$

hvis elementer uden for diagonalen er lig den oprindelige korrelationsmatrix $\underline{\underline{R}}$'s elementer. Denne matrix er stadig symmetrisk, men ikke nødvendigvis længere positiv (semi)definit. Da den er et skøn over en sådan, forudsætter vi imidlertid, at den stadig er det.

Uafhængigt af, hvorledes kommunaliteterne er estimeret, benævnes den resulterende "korrelationsmatrix" $\underline{\underline{V}}$. $\underline{\underline{V}}$ kan således f. eks. være den ovenfor nævnte.

Vi benævner $\underline{\underline{V}}$'s egenværdier og tilhørende normerede, ortogonale egenvektorer

$$\lambda_1 \geq \cdots \geq \lambda_k ,$$

henholdsvis

$$\underline{p}_1 , \cdots , \underline{p}_k .$$

Sætter vi

$$\underline{\underline{P}} = (\underline{p}_1, \cdots, \underline{p}_k) ,$$

har vi ifølge sætning 1.10 p. , at

$$\underline{\underline{P}}' \underline{\underline{V}} \underline{\underline{P}} = \underline{\underline{\Lambda}} = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \lambda_k \end{pmatrix} .$$

Da $\underline{\underline{P}}$ er ortogonal, fås

$$\underline{\underline{V}} = \underline{\underline{P}} \, \underline{\underline{\Lambda}} \, \underline{\underline{P}}' = (\underline{\underline{P}} \, \underline{\underline{\Lambda}}^{\frac{1}{2}}) (\underline{\underline{P}} \, \underline{\underline{\Lambda}}^{\frac{1}{2}})' ,$$

hvor

$$\underline{\underline{\Lambda}}^{\frac{1}{2}} = \begin{pmatrix} \sqrt{\lambda}_1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \sqrt{\lambda}_k \end{pmatrix} .$$

Vi definerer nu

$$\underline{\underline{\Lambda}}_*^{\frac{1}{2}} = \begin{pmatrix} \sqrt{\lambda}_1 & \cdots & 0 \\ \vdots & & \vdots \\ & & \sqrt{\lambda}_m \\ & & \vdots \\ 0 & \cdots & 0 \end{pmatrix} .$$

d.v.s.: $\underline{\underline{\Lambda}}_*^{\frac{1}{2}}$ består af de m første søjler i $\underline{\underline{\Lambda}}^{\frac{1}{2}}$ svarende til de m største egenværdier. Vi ser da

$$(\underline{\underline{P}}\ \underline{\underline{\Lambda}}_*^{\frac{1}{2}})(\underline{\underline{P}}\ \underline{\underline{\Lambda}}_*^{\frac{1}{2}})' = \underline{\underline{P}}\ \underline{\underline{\Lambda}}_*^{\frac{1}{2}}\ \underline{\underline{\Lambda}}_*^{\frac{1}{2}}{}'\underline{\underline{P}}'$$

$$= \underline{\underline{P}} \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_m \end{pmatrix} \underline{\underline{P}}'$$

$$\simeq \underline{\underline{V}} ,$$

jvf. de analoge betragtninger p. 8.7.

Da $\underline{\underline{V}}$ var et skøn over $\underline{\underline{A}}\ \underline{\underline{A}}'$, har vi derfor

$$\underline{\underline{A}}\ \underline{\underline{A}}' \simeq (\underline{\underline{P}}\ \underline{\underline{\Lambda}}_*^{\frac{1}{2}})(\underline{\underline{P}}\ \underline{\underline{\Lambda}}_*^{\frac{1}{2}}) ,$$

hvorfor det vil være naturligt at vælge $\underline{\underline{P}}\ \underline{\underline{\Lambda}}_*^{\frac{1}{2}}$ som skøn over $\underline{\underline{A}}$. Denne løsning kaldes "principal faktor"-løsningen til vort estimationsproblem.

Vi samler overvejelserne i følgende

Sætning 8.7 Vi betragter faktormodellen $\underline{X} = \underline{\underline{A}}\ \underline{F} + \underline{G}$, hvor $\underline{X}$ er k-dimensional og $\underline{F}$ m-dimensional. $\underline{X}$'s korrelationsmatrix betegnes $\underline{\underline{R}}$, og $\underline{\underline{V}}$ er den matrix, der fremkommer ved at erstatte 1-tallene i $\underline{\underline{R}}$'s diagonal med estimater over kommunaliteterne. Disse skal vælges i intervallet $[r^2,1]$, hvor r^2 er den multiple korrelationskoefficient mellem den relevante variabel og de resterende. Sædvanligt vælges enten r^2 eller 1. Principal faktorløsningen til estimationsproblemet er da

$$\underline{\underline{P}}\ \underline{\underline{\Lambda}}_*^{\frac{1}{2}} = (\sqrt{\lambda}_1\ \underline{p}_1, \cdots, \sqrt{\lambda}_m\ \underline{p}_m) ,$$

hvor λ_i, $i = 1, \cdots, m$, er de m største egenvektorer til $\underline{\underline{V}}$, og hvor $\underline{p}_i$, $i = 1, \cdots, m$, er de tilsvarende normerede egenvektorer.

Bemærkning Det forudsættes i sætningen, at antallet af faktorer m er kendt. Hvis dette ikke er tilfældet, er det en udbredt praksis netop at medtage alle, der svarer til egenværdier > 1. Andre foreslår, at man nøjes med en to à tre stykker, fordi det som regel vil være øvre grænse for, hvor mange man kan give en rimelig tolkning af (sic!).

8.3.3 Faktor rotation

Vi betragter igen udtrykket

$$\underline{\underline{A}}\ \underline{\underline{A}}' \simeq (\underline{\underline{P}}\ \underline{\underline{\Lambda}}_*^{\frac{1}{2}})(\underline{\underline{P}}\ \underline{\underline{\Lambda}}_*^{\frac{1}{2}})' .$$

Hvis nu $\underline{\underline{Q}}$ er en vilkårlig m × m ortogonal matrix, d.v.s.: $\underline{\underline{Q}}\ \underline{\underline{Q}}' = \underline{\underline{I}}$, har vi

$$\begin{aligned}(\underline{\underline{P}}\ \underline{\underline{\Lambda}}_*^{\frac{1}{2}}\ \underline{\underline{Q}})(\underline{\underline{P}}\ \underline{\underline{\Lambda}}_*^{\frac{1}{2}}\ \underline{\underline{Q}})' &= (\underline{\underline{P}}\ \underline{\underline{\Lambda}}_*^{\frac{1}{2}})\underline{\underline{Q}}\ \underline{\underline{Q}}'(\underline{\underline{P}}\ \underline{\underline{\Lambda}}_*^{\frac{1}{2}})' \\ &= (\underline{\underline{P}}\ \underline{\underline{\Lambda}}_*^{\frac{1}{2}})(\underline{\underline{P}}\ \underline{\underline{\Lambda}}_*^{\frac{1}{2}}) \\ &\simeq \underline{\underline{A}}\ \underline{\underline{A}}' .\end{aligned}$$

Dette indebærer altså, at vi kan få vilkårligt mange estimater for $\underline{\underline{A}}$-matricen ved at multiplicere principal faktor løsningen med en ortogonal matrix.

Problemet er så blot, hvorledes man hensigtmæssigt vælger $\underline{\underline{Q}}$-matricen. Hovedprincippet er, at man ønsker, at $\underline{\underline{A}}$-matricen bliver "simpel" (uden at komme nærmere ind på, hvad dette så end skal betyde).

Et af de mest benyttede kriterier er det af Kaiser introducerede Varimax kriterium. Det tilsiger, at man skal vælge $\underline{\underline{Q}}$ således, at størrelsen

$$\sum_j m\left\{\sum_i \left(\frac{a_{ij}^2}{h_i^2}\right)^2 - \frac{1}{m}\left[\sum_i \left(\frac{a_{ij}^2}{h_i^2}\right)^2\right]^2\right\}$$

maksimaliseres. Det ses, at udtrykket er den empiriske varians af leddene a_{ij}^2/h_i^2. En maksimalisering vil derfor indebære, at mange af a_{ij}'erne bliver 0 (ca.), og mange bliver store. Og dette svarer jo netop til en simpel struktur, som vil være let at tolke.

Et andet rotationsprincip er det såkaldte quartimax-princip. Sagt med ord tilstræbes det med dette princip, at rækkerne i faktormatricen gøres simple, i.e. at de enkelte variable får en simpel sammenhæng med faktorerne.

I modsætning hertil søger Varimax-kriteriet at gøre søjlerne simple svarende til, at man ønsker let tolkelige faktorer.

Inden vi fortsætter med teorien, giver vi et eksempel.

Eksempel 8.3 Vi vil nu forsøge at lave en faktoranalyse på de i eksempel 8.1 anførte data.

Vi bestemmer først korrelationsmatricen. Ud fra estimatet på dispersionsmatricen p. 8.11 finder vi

$$\hat{\underline{\underline{R}}} = \begin{bmatrix} 1.000 & 0.580 & 0.201 & 0.911 & 0.283 & 0.287 & -0.533 \\ 0.580 & 1.000 & 0.364 & 0.834 & 0.166 & 0.261 & -0.609 \\ 0.201 & 0.364 & 1.000 & 0.439 & -0.704 & -0.681 & -0.649 \\ 0.911 & 0.834 & 0.439 & 1.000 & 0.163 & 0.202 & -0.676 \\ 0.283 & 0.166 & -0.704 & 0.163 & 1.000 & 0.990 & 0.427 \\ 0.287 & 0.261 & -0.681 & 0.202 & 0.990 & 1.000 & 0.357 \\ -0.533 & -0.609 & -0.649 & -0.676 & 0.427 & 0.357 & 1.000 \end{bmatrix}$$

Fuldstændigt i analogi med fremgangsmåden i eksempel 8.1 bestemmes dernæst egenværdier og -vektorer for $\hat{\underline{\underline{R}}}$. Vi finder

Egenværdi $\hat{\lambda}_i, i=1,\cdots,7$	Procentdel af total varians	Kumuleret procentdel af total varians
3.3946	48.495	48.495
2.8055	40.078	88.573
0.4373	6.247	94.820
0.2779	3.971	98.791
0.0810	1.157	99.948
0.0034	0.049	99.996
0.0003	0.004	100.000

De tilsvarende egenvektorers koordinater er vist i nedenstående tabel.

Variabel	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$	$\hat{p}_4$	$\hat{p}_5$	$\hat{p}_6$	$\hat{p}_7$
X_1	0.405	-0.293	-0.667	0.089	-0.227	0.410	-0.278
X_2	0.432	-0.222	0.698	-0.034	-0.437	0.144	-0.254
X_3	0.385	0.356	0.148	0.628	0.512	0.188	-0.108
X_4	0.494	-0.232	-0.119	0.210	-0.105	-0.588	5.536
X_5	-0.128	-0.575	0.209	0.111	0.389	-0.423	-0.556
X_6	-0.097	-0.580	0.174	-0.006	0.355	0.500	0.498
X_7	-0.481	-0.130	0.018	0.735	-0.455	0.033	0.049

Vi antager, at antallet af faktorer er 2 (antagelsen er ikke begrundet ved dybere overvejelser over strukturen i problemet. Tallet 2 er valgt, fordi der kun er 2 egenværdier større end 1).

Ifølge sætning 8.7 er den estimerede, principale faktor-løsning til problemet $(\sqrt{\hat{\lambda}}_1 \hat{\underline{p}}_1, \sqrt{\hat{\lambda}}_2 \hat{\underline{p}}_2)$, hvor

$$\begin{pmatrix} \sqrt{\hat{\lambda}}_1 \, \hat{\underline{p}}_1' \\ \sqrt{\hat{\lambda}}_2 \, \hat{\underline{p}}_2' \end{pmatrix} = \begin{pmatrix} 0.747 & 0.795 & 0.710 & 0.910 & -0.235 & -0.178 & -0.886 \\ -0.491 & -0.373 & 0.596 & -0.389 & -0.963 & -0.971 & -0.218 \end{pmatrix}.$$

Vi kan nu også finde skøn over kommunaliteten på hver af de variable.

Vi finder eksempelvis

$$\hat{h}_7^2 = (-0.886)^2 + 0.218^2 = 0.833$$

Vektoren af skønnede kommunaliteter er

$$\hat{\underline{h}}^{2\prime} = [0.798 \quad 0.771 \quad 0.860 \quad 0.979 \quad 0.983 \quad 0.976 \quad 0.833],$$

og vi ser, at f. eks. variationen i variabel 4 (længden af den længste diagonal) for 97.9% vedkommende kan beskrives ved variationen i de to faktorer.

Omvendt angiver størrelserne $\hat{\delta}_j = 1-\hat{h}_j^2$ ("uniqueness"-værdien) den brøkdel af X_j's varians, der ikke forklares af de to fælles faktorer, men som hidrører fra den j'te unikke faktor G_j (jvf. p. 8.22). Vi finder

$$\underline{\delta}' = [0.202 \quad 0.229 \quad 0.140 \quad 0.021 \quad 0.017 \quad 0.024 \quad 0.167].$$

Et lidt mere kvalificeret mål for de to faktorers evne til at beskrive variationen i materialet fås ved at søge at genberegne korrelationsmatricen ud fra faktorerne alene.

Vi bestemmer derfor den såkaldte residual-korrelationsmatrix

$$\hat{\underline{\underline{Z}}} = \hat{\underline{\underline{R}}} - \hat{\underline{\underline{A}}}\,\hat{\underline{\underline{A}}}' \ ,$$

og får som mål for faktorernes evne til at beskrive den oprindelige variabilitet i materialet

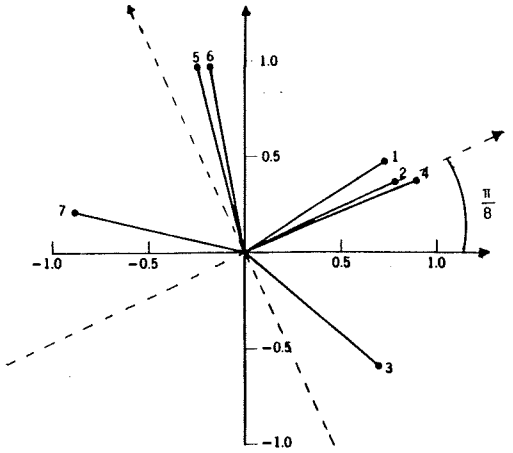
$$\hat{\underline{\underline{Z}}} = \begin{bmatrix} 0.202 & -0.196 & -0.037 & 0.041 & -0.914 & -0.057 & 0.021 \\ -0.196 & 0.229 & 0.071 & -0.035 & -0.006 & 0.041 & 0.015 \\ -0.037 & 0.021 & 0.140 & 0.024 & 0.037 & 0.025 & 0.111 \\ 0.041 & -0.035 & 0.024 & 0.021 & 0.002 & -0.013 & 0.046 \\ -0.014 & -0.006 & 0.037 & 0.002 & 0.017 & 0.012 & 0.009 \\ -0.057 & 0.041 & 0.025 & -0.013 & 0.012 & 0.024 & -0.013 \\ 0.021 & 0.015 & 0.111 & 0.046 & 0.009 & -0.013 & 0.167 \end{bmatrix} .$$

Jo mere $\hat{\underline{\underline{Z}}}$ afviger fra $\underline{\underline{0}}$-matricen, jo dårligere beskriver faktorerne det oprindelige materiale.

Den væsentlige forskel på den analyse, der blev foretaget i eksempel 8.1 og her, er - bortset fra at vi der arbejdede på dispersionsmatricen, hvor vi her arbejder med korrelationsmatricen - at vi har multipliceret faktorerne med kvadratroden af den til hver faktor svarende egenværdi. Derved bliver længden af en faktor proportional med den del af den totale varians, som den forklarer.

Vi vil nu se, om vi kan opnå lettere tolkelige faktorer ved at rotere disse.

Vi afbilder først faktorvægtene (angivet p. 8.30) $\hat{a}_{ij}$ i et todimensionalt koordinatsystem. Vi finder



Det ses, at de fleste variable har såvel første som anden koordinat jævnt store.

Det synes muligt at opnå en simplere struktur ved at rotere koordinatsystemet ca. $\frac{\pi}{8}$ (= $22\frac{1}{2}^{\circ}$) mod urviseren.

Dette svarer til multiplikation med matricen

$$\begin{pmatrix} \cos\frac{\pi}{8} & -\sin\frac{\pi}{8} \\ \sin\frac{\pi}{8} & \cos\frac{\pi}{8} \end{pmatrix} = \begin{pmatrix} 0.9239 & -0.3827 \\ 0.3827 & 0.9239 \end{pmatrix},$$

jvf. afsnit 1.4.1.

De nye faktorer - eller rettere faktorvægte - bliver da

$$\begin{bmatrix} 0.747 & 0.491 \\ 0.795 & 0.373 \\ 0.710 & -0.596 \\ 0.910 & 0.389 \\ -0.235 & 0.963 \\ -0.178 & 0.971 \\ -0.886 & 0.218 \end{bmatrix} \begin{bmatrix} 0.9239 & -0.3827 \\ 0.3827 & 0.9239 \end{bmatrix} = \begin{bmatrix} 0.878 & 0.168 \\ 0.877 & 0.040 \\ 0.428 & -0.822 \\ 0.990 & 0.011 \\ 0.151 & 0.980 \\ 0.207 & 0.965 \\ -0.735 & 0.540 \end{bmatrix}.$$

Disse nye faktorvægte er simplere end de oprindelige i den forstand, at der optræder flere størrelser i nærheden af 1 og flere i nærheden af 0. Vi skal senere se, at denne visuelt fundne løsning ligger ganske nær Varimax-løsningen.

□

Foruden Varimax-princippet findes som nævnt en lang række andre metoder til ortogonal rotation af faktorer, og det ligger uden for denne fremstillings rammer at komme ind på beskrivelsen af disse. Den interesserede læser må henvises til litteraturen (e.g. Harman (1967) eller Cattell (1965)).

Der findes også en række rotationsmetoder, hvor kravet om ortogonalitet ikke opretholdes. Disse rotationsmetoder kaldes "oblique rotation". Filosofien bag disse er, at faktorer ikke nødvendigvis behøver at være uafhængige, men godt kan være korrelerede. En anvendelse af disse metoder kræver dog et yderligere godt kendskab til emnet. Der kan igen henvises til Harman (1967) og Cattell (1965).

8.3.4 Beregning af faktorværdier (factor scores)

Hvis vi i ovenstående eksempel 8.3 ønsker at lave et diagram analogt til det p. 8.13 anførte, må man beregne faktorværdierne (scores) for de enkelte æsker. Dette er en anelse mere kompliceret, end det var ved den principale komponentanalyse, hvor vi blot skulle bestemme værdierne af de principale komponenter på de forskellige akser. Grunden til, at vi ikke blot kan foretage den analoge operation, er tilstedeværelsen af de specifikke faktorer $\underline{G}$.

Vi har modellen (jvf. p. 8.21)

$$\underline{X} = \underline{\underline{A}}\,\underline{F} + \underline{G} \ ,$$

hvor

$$D(\underline{F}) = \underline{\underline{I}}$$

$$D(\underline{G}) = \underline{\underline{\Delta}} \ ,$$

og hvor $\underline{F}$ og $\underline{G}$ er ukorrelerede.

Derfor er

$$D\begin{pmatrix} \underline{X} \\ \underline{F} \end{pmatrix} = \begin{pmatrix} \underline{\underline{A}}\,\underline{\underline{A}}' + \underline{\underline{\Delta}} & \underline{\underline{A}} \\ \underline{\underline{A}}' & \underline{\underline{I}} \end{pmatrix} .$$

Da som tidligere nævnt

$$\mathrm{Cov}(X_i, F_j) = a_{ij} \ ,$$

følger, at matricerne uden for diagonalen netop er $\underline{\underline{A}}$-matricen, respektive dens transponerede.

Skønnet over denne dispersionsmatrix er

$$\begin{bmatrix} \hat{\underline{\underline{A}}}\,\hat{\underline{\underline{A}}}' + \hat{\underline{\underline{\Delta}}} & \hat{\underline{\underline{A}}} \\ \hat{\underline{\underline{A}}}' & \underline{\underline{I}} \end{bmatrix} .$$

Den betingede fordeling af $\underline{F}$ for givet $\underline{X}$ har - såfremt de underliggende fordelinger er normale - som middelværdi

$$\underline{\mu}_F + \underline{\underline{A}}'(\underline{\underline{A}}\,\underline{\underline{A}}' + \underline{\underline{\Delta}})^{-1}(\underline{x} - \underline{\mu}_x)$$

(jvf. afsnit 2.2.3).

Da vi foretager vore beregninger på de standardiserede x-værdier, er det rimeligt at antage, at $\underline{\mu}_x = \underline{0}$. Niveauet for faktorskalaerne er arbitrært, men det er sædvane også at sætte det lig 0, således at vi får udtrykket

$$\underline{\underline{A}}'(\underline{\underline{A}}\,\underline{\underline{A}}' + \underline{\underline{\Delta}})^{-1}\underline{x}$$

for den betingede middelværdi af $\underline{F}$.

Som estimat af den i'te måling af $\underline{X}_i$'s faktorværdi har vi da

$$\hat{\underline{F}}_i = \hat{\underline{\underline{A}}}'(\hat{\underline{\underline{A}}}\,\hat{\underline{\underline{A}}}' + \hat{\underline{\underline{\Delta}}})^{-1}\,\underline{X}_i \quad . \tag{1}$$

Nu vil A-matricen ofte have et stort antal rækker, hvorfor vi nødsages til at invertere en ret stor matrix. Dette kan omgås ved hjælp af følgende identitet

$$(\underline{\underline{A}}\,\underline{\underline{A}}' + \underline{\underline{\Delta}})^{-1}\,\underline{\underline{A}} = \underline{\underline{\Delta}}^{-1}\,\underline{\underline{A}}(\underline{\underline{I}} + \underline{\underline{A}}'\,\underline{\underline{\Delta}}^{-1}\,\underline{\underline{A}})^{-1} \ ,$$

som giver

$$\hat{\underline{F}}_i = (\underline{\underline{I}} + \hat{\underline{\underline{A}}}'\,\hat{\underline{\underline{\Delta}}}^{-1}\,\hat{\underline{\underline{A}}})^{-1}\,\hat{\underline{\underline{A}}}'\,\hat{\underline{\underline{\Delta}}}^{-1}\,\underline{X}_i \quad . \tag{2}$$

Identitetens gyldighed fremgår af følgende relationer

$$(\underline{\underline{A}}\,\underline{\underline{A}}' + \underline{\underline{\Delta}})^{-1}\,\underline{\underline{A}} = \underline{\underline{\Delta}}^{-1}\,\underline{\underline{A}}(\underline{\underline{I}} + \underline{\underline{A}}'\,\underline{\underline{\Delta}}^{-1}\,\underline{\underline{A}})^{-1}$$

$$\leftrightarrow \qquad \underline{\underline{A}} = (\underline{\underline{A}}\,\underline{\underline{A}}' + \underline{\underline{\Delta}})\underline{\underline{\Delta}}^{-1}\,\underline{\underline{A}}(\underline{\underline{I}} + \underline{\underline{A}}'\,\underline{\underline{\Delta}}^{-1}\,\underline{\underline{A}})^{-1}$$

$$= \underline{\underline{A}}(\underline{\underline{A}}'\,\underline{\underline{\Delta}}^{-1}\,\underline{\underline{A}} + \underline{\underline{I}})(\underline{\underline{I}} + \underline{\underline{A}}'\,\underline{\underline{\Delta}}^{-1}\,\underline{\underline{A}})^{-1} \ ,$$

og den sidste relation er jo trivielt opfyldt.

Nu er $\underline{\underline{I}} + \underline{\underline{A}}'\,\underline{\underline{\Delta}}^{-1}\,\underline{\underline{A}}$ en m x m matrix, hvor m er antallet af faktorer, d.v.s. ofte ikke mere end en 2-3-4 stykker, hvorfor inversionsproblemet ikke er overvældende. Derimod er som nævnt $(\underline{\underline{A}}\,\underline{\underline{A}}' + \underline{\underline{\Delta}})$ en k x k matrix, hvor k er antallet af variable, d.v.s. oftest langt større end m.

Hvis k kun er moderat stor, kan man dog godt anvende det første udtryk for F_i direkte. Her bør man så benytte, at

$$\underline{\underline{R}} = \underline{\underline{A}}\,\underline{\underline{A}}' + \underline{\underline{\Delta}}$$

(jvf. p. 8.23). Dette giver det med (1) ækvivalente udtryk

$$\hat{\underline{F}}_i = \hat{\underline{\underline{A}}}' \, \hat{\underline{\underline{R}}}^{-1} \, \underline{X} \qquad (3)$$

Det må slutteligen præciseres, at der findes en række andre metoder til bestemmelse af faktorværdier, se f. eks. Harman (1967) eller Morrison (1967). Det må i øvrigt bemærkes, at problemet er ret svagt behandlet i den overvejende del af litteraturen. Det skyldes væsentligst, at dette problem ikke har haft den store interesse for psykologer og sociologer, som i mange år har været de væsentlige brugere af faktoranalysen. I en række teknisk-naturvidenskabelige (og sociologiske) anvendelser er man imidlertid ofte interesseret i at få klassificeret enkeltmålinger efter størrelsen af faktorværdier. Dette skal vi se en anvendelse af i afsnit 8.3.5.

Vi vil nu illustrere beregningen af faktorværdier (factor scores) på vort kasseeksempel.

Eksempel 8.4 I eksempel 8.3, p. 8.28, fandt vi en roteret faktorløsning med 2 faktorer. De roterede faktorvægte var

$$\hat{\underline{\underline{A}}} = \begin{bmatrix} 0.878 & 0.168 \\ 0.877 & 0.040 \\ 0.428 & -0.828 \\ 0.990 & 0.011 \\ 0.151 & 0.980 \\ 0.207 & 0.965 \\ -0.735 & 0.540 \end{bmatrix} .$$

For at bestemme faktorværdierne for de enkelte kasser må vi først finde kommunaliteterne og uniqueness-værdierne. Vi finder

j	1	2	3	4	5	6	7
$\hat{h}_j^2$	0.7991	0.7707	0.8589	0.9802	0.9832	0.9741	0.8318
$\hat{\delta}_j$	0.2009	0.2293	0.1411	0.0198	0.0168	0.0259	0.1682
$1/\hat{\delta}_j$	4.9776	4.3611	7.0872	50.5051	59.5238	38.6100	5.9453

Her er (jvf. p. 8.23)

$$\hat{h}_j^2 = \hat{a}_{j1}^2 + \hat{a}_{j2}^2 = 1 - \hat{\delta}_j \ .$$

Vi bemærker, at de her angivne kommunaliteter er lig dem, vi fandt p. 8.30 for de uroterede faktorer. Dette er alment gyldigt og kan anvendes som et check ved beregningen af de roterede faktorer.

Idet vi har

$$\hat{\underline{\underline{\Delta}}} = \text{diag}(\hat{\delta}_j) \ ,$$

d.v.s.

$$\hat{\underline{\underline{\Delta}}}^{-1} = \text{diag}\left(\frac{1}{\hat{\delta}_j}\right) \ ,$$

bliver

$$(\underline{\underline{I}} + \hat{\underline{\underline{A}}}'\hat{\underline{\underline{\Delta}}}^{-1}\hat{\underline{\underline{A}}})^{-1}\hat{\underline{\underline{A}}}'\hat{\underline{\underline{\Delta}}}^{-1}$$

$$= \begin{bmatrix} 0.0669 & 0.0597 & 0.0593 & 0.7839 & 0.0244 & 0.0510 & -0.0750 \\ -0.0002 & -0.0059 & 0.0655 & -0.0943 & 0.5770 & 0.3641 & 0.0415 \end{bmatrix}$$

Formel (3) forudsætter, at de variable X er standardiserede. Vi må derfor først bestemme middelværdi og spredning for hver af de 7 variable. Disse er

j	1	2	3	4	5	6	7
$\bar{X}_{\cdot j}$	7.1000	4.7730	2.3488	9.1338	5.4582	7.1674	2.3462
s_j	2.3238	2.4178	1.6656	3.0178	3.2733	4.5581	1.6105

De standardiserede værdier for eksempelvis den første æske bliver derfor

$$\underline{z} = (-1.4373,\ -0.4603,\ -1.0860,\ -1.2787,\ 1.3167,\ 1.4422,\ 1.5124)' ,$$

hvor f. eks. den anden værdi fremkommer som

$$z_2 = \frac{3.660-4.773}{2.4178} = -0.4603 .$$

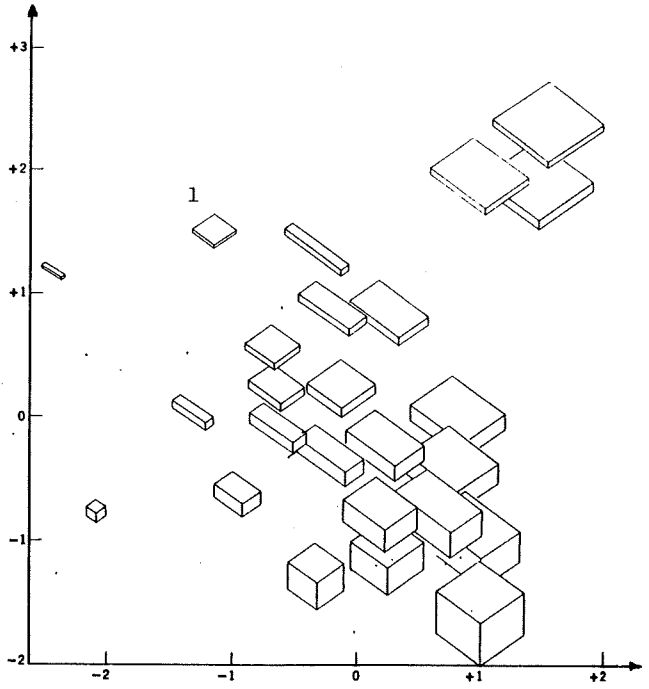
Vi finder nu let faktorværdierne svarende til den første æske som

$$\hat{\underline{F}}^1 = (\underline{\underline{I}} + \hat{\underline{\underline{A}}}'\hat{\underline{\underline{\Delta}}}^{-1}\hat{\underline{\underline{A}}})^{-1}\hat{\underline{\underline{A}}}'\hat{\underline{\underline{\Delta}}}^{-1}\underline{z} = \begin{pmatrix} -1.20 \\ 1.40 \end{pmatrix} .$$

De øvrige bestemmes selvsagt analogt.

I nedenstående figur er i et 2-dimensionalt koordinatsystem indtegnet de 25 æsker således, at hver æske er anbragt på de koordinater, der svarer til dens faktorværdier (jvf. p. 8.13).

Vi ser (jvf. eksempel 8.3), at de to faktorer beskriver "tykkelse" og "størrelse". Vi bemærker dog, at der er byttet om på "vigtigheden" af de to begreber i forhold til eksempel 8.1.



8.3.5 Et case-study

I dette afsnit anfører vi en artikel (fra Geografisk Tidsskrift, 1972, 49-56) af Poul Ove Pedersen og Peter Rasmussen vedrørende den indre differentiering af (og mellem) danske provinsbyer (Århus, Odense og Ålborg). Analysen foregår ved hjælp af en faktoranalyse, og det illustreres, hvorledes man kan bruge faktoranalysen - eller mere præcist faktorværdierne - ved en klassifikation af de tre byer. Notation og begreber svarer i øvrigt nøje til de i denne fremstilling anvendte.

Danske provinsbyers indre differentiering og differentieringen mellem danske provinsbyer

Af Poul Ove Pedersen og Peter Rasmussen

Pedersen, P. O. & Rasmussen, P., 1973: Danske provinsbyers indre differentiering og differentieringen mellem danske provinsbyer. Geografisk Tidsskrift, 72, 49–56. København, september 30., 1973.

This paper analyses the inner differentiation in the three largest provincial towns in Denmark by means of a factor analysis of 25 variables characterizing the population and the housing in 40 zones, 14 in Århus (187,000 inh.), 14 in Odense (133,000 inh.), and 12 in Ålborg (123,000 inh.). The paper especially focuses on the differences in inner differentiation between the three towns.

Civilingeniør P. O. Pedersen, Institute for Road Construction, Traffic Engineering and Townplanning. Technical University. Lyngby DK 2800. Civilingeniør P. Rasmussen, Stadt Stuttgart Stadtplanungsamt. Stuttgart West D 7000.

Indledning

Faktoranalyser af befolkningens geografiske fordeling er efterhånden lavet for mange store byer. Disse analyser viser en række slående ligheder mellem de geografiske strukturer af den vestlige verdens storbyer. I næsten alle de analyserede byer har størstedelen af den indre differentiering ($^2/_3$ eller mere af den samlede varians mellem

Tabel 1. Faktorvægtene for de tre første principale faktorer. Faktorerne er Varimaxroterede.

	Faktor 1 Familiestatus	Faktor 2 Socio-økonomisk status	Faktor 3 Byspecialisering	Kommunaliteter
1. Pct. af befolkningen i aldersgruppen 0 – 14 år	0,37	– 0,09	– 0,04	0,35
2. " " " " " 15 – 24 år	–0,35	0,24	– 0,62	0,56
3. " " " " " 25 – 64 år	–0,38	– 0,13	0,67	0,61
4. " " " " " 64 år	–0,92	0,07	0,02	0,86
5. Pct. af befolkningen, der er kvinder	–0,85	0,12	– 0,15	0,76
6. Pct. af kvinder, der er gifte	0,56	– 0,12	0,64	0,75
7. Pct. af befolkningen ernæret ved landbrug	0,76	0,20	– 0,14	0,64
8. " " " " " håndværk og industri	0,53	– 0,66	0,49	0,95
9. " " " " " handel	0,10	0,71	0,28	0,60
10. " " " " " transport	0,07	– 0,43	– 0,67	0,64
11. " " " " " administration og lib.erhv.	0,27	0,80	– 0,30	0,81
12. " " " " " formue, rente	–0,91	0,09	– 0,28	0,92
13. Pct. af befolkningen, der er erhvervsmæssigt beskæftiget	–0,70	– 0,26	0,39	0,79
14. Pct. af de beskæftigede, der er selvstændige	–0,30	0,51	0,16	0,78
15. " " " " " " funktionærer	0,25	0,80	– 0,08	0,71
16. " " " " " " arbejdere	–0,12	– 0,93	0,02	0,88
17. Pct. af kvinder der er erhvervsmæssigt beskæftiget	–0,87	– 0,13	– 0,00	0,78
18. Pct. af alle lejligheder i landbrugsejendomme	0,19	0,20	0,16	0,64
19. " " " " " enfamiliehuse	0,28	0,42	0,72	0,92
20. " " " " " tofamiliehuse	0,02	0,21	0,69	0,52
21. " " " " " større beboelsesejendomme m.m.	–0,77	– 0,44	– 0,38	0,93
22. Gennemsnitligt antal værelser pr. lejlighed	0,71	0,58	0,32	0,95
23. " " personer pr. husstand	0,98	– 0,06	0,00	0,96
24. " " værelser pr. person	–0,43	0,77	0,29	0,91
25. " " husstande pr. lejlighed	0,34	0,69	– 0,17	0,61

zonerne i byen) kunnet beskrives ved hjælp af de samme tre faktorer (se f.eks. sammenstillingen i D.W.G. Timms (1971) tabel 2.3.):

– en faktor, der normalt kaldes familiestruktur eller livscyklusstatus, og som beskriver befolkningens demografiske karakteristika. Den skelner mellem byens perifere områder med mange unge husstande og små børn og de indre bydele med aldrende befolkning, og udviser derfor normalt et geografisk mønster af ringe omkring bymidten;

– en faktor, der kaldes socio-økonomisk status, og som beskriver befolkningens erhverv og beskæftigelsesmæssige status. Den skelner mellem områder med overvejende arbejderbefolkning og områder med overvejende funktionærbefolkning, og følger ofte et sektormønster med sektorer der stråler ud fra bymidten, og endelig

– en faktor, der i byer med store racemæssige, religiøse eller sproglige mindretal ofte kaldes segregation, fordi den udskiller de områder der er domineret af disse mindretal. I byer hvor sådanne mindretal er små udskiller den ofte i stedet for områder med mange tilflyttere og mange ugifte unge. Rees (1970) fandt således for Chicago en faktor han kalder Immigrant and Catholic, Sweetser (1965 a og b) fandt for Helsingfors en faktor han kalder progeniture, fordi den udskiller med mange 15–24 årige, d.v.s. netop den aldersgruppe der foretager flest vandringer, og Pedersen (1967) fandt for København en faktor han kalder vækst og mobilitet.

Selv om ligheden mellem byernes indre differentiering således er slående, så er der naturligvis også forskelle; og vi ved fra utallige analyser af andre aspekter ved byer end den indre differentiering, at byerne afviger fra hinanden på andre punkter, nogle vokser hurtigt, medens andre stagnerer, nogle er rige, medens andre er fattige; og nogle er bare oplandsbyer, medens andre desuden har specialiseret sig som f.eks. industricentrer, administrationscentrer eller transportknudepunkter.

I dette notat skal vi kæde disse to typer af analyser sammen, og forsøge at vise, hvorledes byernes forskellige rolle i bysystemet påvirker deres indre differentiering.

I sin artikel: Cities as Systems within Systems of Cities nærmede Berry (1964) sig dette problem, men han behandlede den enkelte bys interne system og det overordnede system af byer som uafhængige af hinanden, og det er netop denne afhængighed, der er emnet for denne artikel.

Metoden: En faktoranalyse

Vort udgangspunkt er en faktoranalyse af den indre differentiering af Danmarks tre største provinsbyer, Århus, Odense og Ålborg. Disse tre byer er valgt til analysen, fordi det er de eneste danske provinsbyer, for hvilke der foreligger detaljerede folketællingsoplysninger for et tilstrækkeligt antal zoner til at muliggøre analyser af den interne differentiering. Desuden er de tre byer af samme

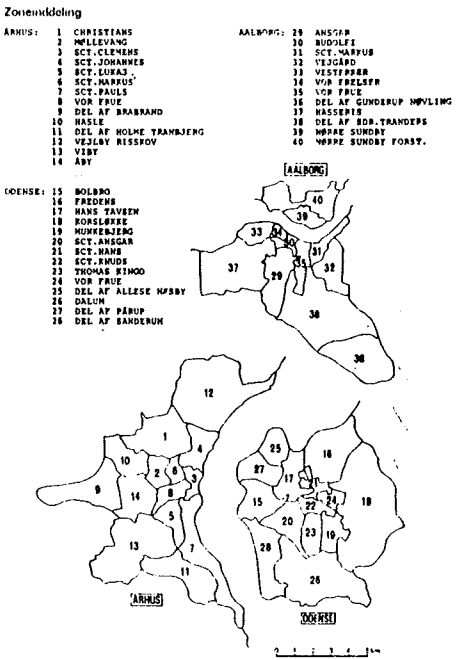


Fig. 1. Den anvendte zoneinddeling af Århus, Odense og Ålborg. *Fig. 1. The 40 zones applied in the analysis, distributed on the three towns of Århus, Odense, and Ålborg.*

størrelsesorden og de eneste danske provinsbyer, der indiskutabelt er overordnede regionale centrer (Illeris og Pedersen, 1968).

For disse tre byer har vi analyseret en datamatrix med 25 variable og 40 zoner, hvoraf 14 er fra Århus, 14 er fra Odense og 12 er fra Ålborg. De variable karakteriserer befolkningen og boligmassen i de 40 zoner. Det nøjagtige valg af variable fremgår af tabel 1, og zoneinddelingen af figur 1.

For på een gang at kunne analysere variationen mellem zonerne i hver by og variationen mellem de tre byer er samtlige 40 zoner inkluderet i den samme analyse. Denne metode har tidligere været anvendt af Carl-Gunnar Janson (1971) i en analyse af 12 svenske byer.

Vor datamatrix kan afbildes som 40 punkter i et koordinatsystem med 25 akser. Da de 25 variable er indbyrdes korrelerede, vil det 25-dimensionale koordinatsystem (variabelrummet) ikke være retvinklet, men det kan ved hjælp af en faktoranalyse drejes ind til et retvinklet koordinatsystem (faktorrummet) med færre end 25 akser,

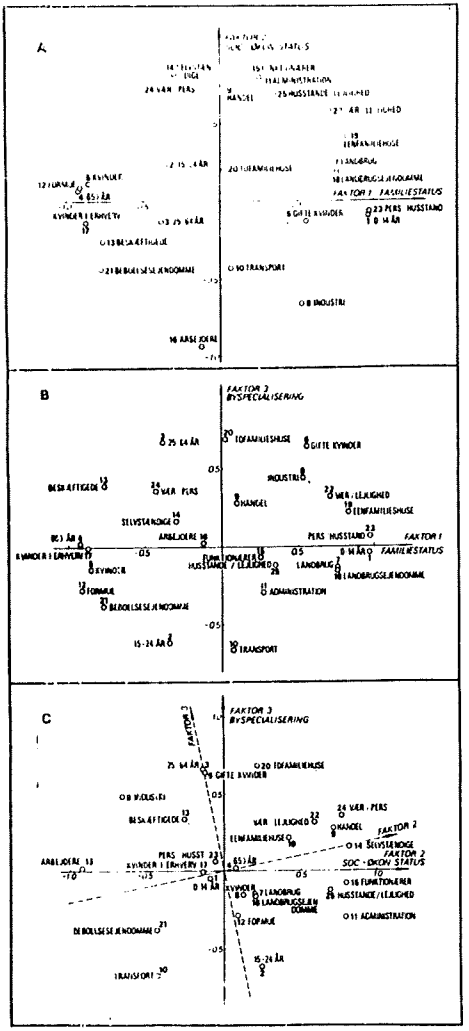


Fig. 2. Diagrammer af sammenhængen mellem faktorvægtene for de tre principale faktorer. Hvert punkt i diagrammerne svarer til en variabel.

Fig. 2. Diagrams showing the interplay between the factor weights for the three principal factors. Each point corresponds to one variable.

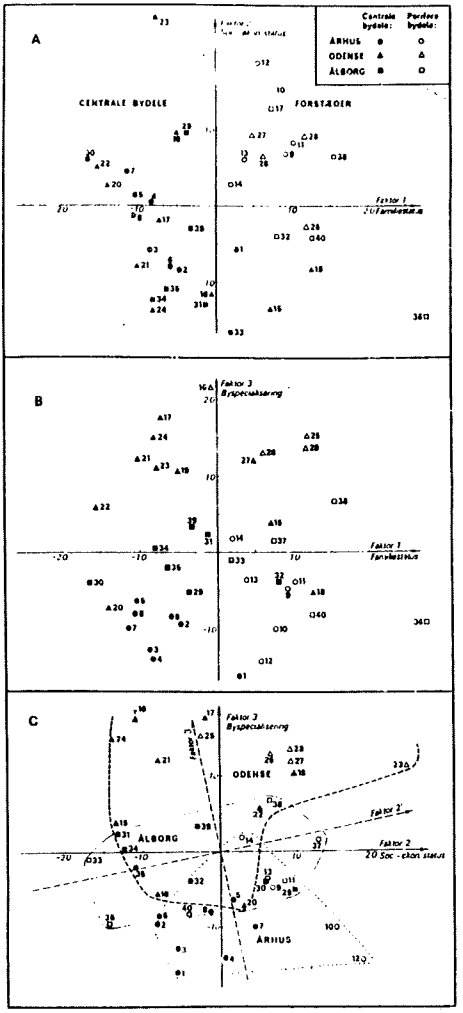


Fig. 3. Diagrammer af sammenhængen mellem faktorværdierne for de tre faktorer. De tre diagrammer viser de tre plane projektioner af det tredimensionale faktorrum. Hvert punkt i diagrammerne svarer til en zone.

Fig. 3. Diagrams showing the interplay between the factor values for the three factors. The diagrams show the three plane projections of the three-dimensional factor space. Each point corresponds to one zone.

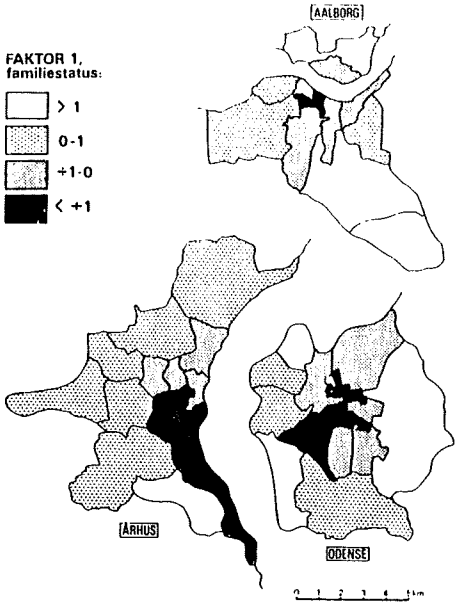


Fig. 4. Kort over faktorværdierne for faktor 1, familiestatus.
Fig. 4. Map showing the factor values for factor 1, family status.

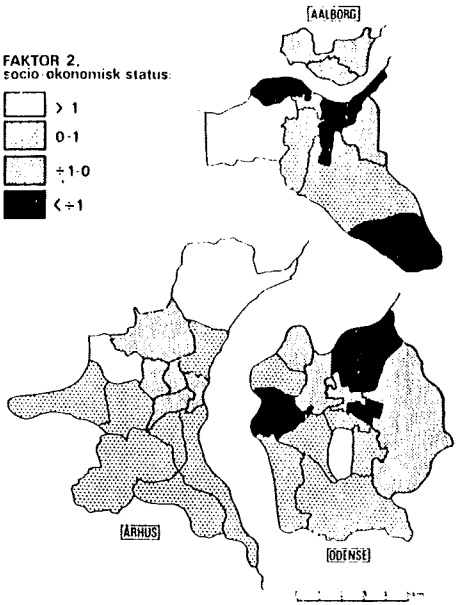


Fig. 5. Kort over faktorværdierne for faktor 2, socio-økonomisk status.
Fig. 5. Map showing the factor values for factor 2, socio-economic status.

d.v.s. at de 25 variable kan reduceres til et mindre antal principale faktorer.

Vi har her taget de tre vigtigste af disse principale faktorer op til nøjere analyse. Disse tre faktorer repræsenterer ialt 77,7 % af den samlede varians i observationsmatricen, nemlig henholdsvis 41,6 %, 23,1 % og 13,0 %. Den fjerde ikke-analyserede faktor repræsenterer 6,1 % af variansen og den femte 3,7 %. Faktoranalyseberegningerne er udført på BMD-program 03M (Biomedical Computer Programs, 1970). For en detaljeret redegørelse for faktoranalysens teori og metode se Harman (1960).

Tolkningen af de principale faktorer

Faktorernes indhold kan fortolkes ved hjælp af faktorvægtene (vist i tabel 1 og figur 2) og faktorværdierne (vist i figur 3–6).

Faktorvægtene er korrelationskoefficienter mellem faktorerne og de 25 oprindelige variable. Faktorvægtene varierer derfor mellem –1 og +1. De variable, der har numerisk høje faktorvægte for en given faktor, er derfor vigtige for forståelsen af den pågældende faktor, medens faktorvægte nær nul er uvæsentlige. Faktorvægtene kaldes tilsammen mønsteret (the factor pattern).

Dette mønster kan afbildes i et koordinatsystem med ligeså mange retvinklede akser som der er principale faktorer, her tre. Diagrammerne i figur 2 viser de to-dimensionale projektioner af dette 3-dimensionale koordinatsystem. De viser tilsammen sammenhængen mellem faktorerne og de variable og også mellem de variable indbyrdes.

Faktorværdierne er koordinaterne til de 40 punkter (zoner) i det tre-dimensionale faktorrum på samme måde som de oprindelige variable er koordinaterne til de 40 punkter i det 25-dimensionale variabelrum. Diagrammerne i figur 3 viser de to-dimensionale projektioner af faktorrummet. Diagrammerne i figur 3 viser derfor også de enkelte zoners positioner i faktorrummet. Figur 4–6 viser den geografiske fordeling af de tre faktorers faktorværdier.

Faktor 1: Familiestatus

Faktor 1 har høje positive faktorvægte for aldersgruppen 0–14 år, for beskæftigelsen i landbruget, for boliger i landbrugsejendomme og eenfamiliehuse, for husstandsstørrelse og for boligstorrelse, og høje negative faktorvægte for aldersgruppen over 64 år, for folk der er ernæret af for-

inue og rente, for erhvervsaktive ialt og erhvervsaktive kvinder samt for boliger i etageejendomme.

Som det fremgår både af figur 3A og 4, har faktoren i alle tre byer høje faktorværdier i de nye perifere bydele og lave faktorværdier i de centrale bydele.

Faktoren svarer nøje til den faktor: familiestatus eller livscyklus status, der i næsten alle analyserede byer har været fundet som en af de 2 vigtigste faktorer.

Faktor 2: Socio-økonomisk status

Faktor 2 har høje faktorvægte for beskæftigelsen i handel og administration, for selvstændige og funktionærer, for antal værelser pr. lejlighed og for antal værelser pr. person. Faktor 2 har også høj positiv faktorvægt for antal husstande pr. lejlighed. Da denne variabel var medtaget som et mål for boligmangelen, burde den være negativt korreleret med faktor 2. Forklaringen på den positive faktorvægt er, at der især forekommer mange husstande pr. lejlighed i de meget store boliger hvor værelser lejes ud, d.v.s. i de relativt velstående kvarterer. Den variable er derfor et dårligt mål for boligmangelen. Faktor 2 har høje negative faktorvægte for beskæftigelsen i industri og håndværk og for arbejdere. Faktoren har i alle tre byer de største værdier i nogle af de perifere zoner. Med lidt god vilje kan faktorens geografiske udbredelse godt tolkes som et sektormønster, således som man har fundet det for den socio-økonomiske statusfaktor i andre større byer; men da antallet af zoner i vores relativt små byer er meget lille, fremtræder sektorerne ikke klart.

Der kan dog ikke være tvivl om, at denne faktor er helt analog med den socio-økonomiske statusfaktor, der i næsten alle analyserede byer er blevet fundet som den anden af de to vigtigste faktorer.

Faktor 3: Byspecialisering

Her som i de fleste andre faktoranalyser begynder fortolkningsproblemerne først med faktor 3. Denne faktor har høje positive faktorvægte for den erhvervsaktive aldersgruppe 25–64 år, for beskæftigelsen i industrien og for boliger i tofamilieshuse, og høje negative faktorvægte for aldersgruppen 15–24 år og for beskæftigelsen i transport, og disse variable giver ikke umiddelbart grundlag for en klar fortolkning af faktoren.

De ovennævnte variable, der har høje faktorvægte for faktor 3, har næsten alle meget lave kommunaliteter, hvilket vil sige at de ikke er særligt godt forklaret ved hjælp af de tre principale faktorer, dette vanskeliggør naturligvis yderligere fortolkningen af faktor 3.

Løsningen på dette tolkningsproblem ligger i figur 3 C, der viser sammenhængen mellem faktorværdierne for faktorerne 2 og 3. Her er zonerne fra de tre byer vist med forskellige signaturer. Det fremgår af figuren, at faktor 3 på få undtagelser nær adskiller zonerne i de tre byer fra hinanden, idet zonerne i Odense har de største faktorværdier, zonerne i Århus har de laveste, og zonerne i Ålborg falder i midten. Drejer vi faktor 2-faktor 3-koordinatsystemet ind til koordinatsystemet faktor 2'-faktor 3' bliver adskillelsen mellem de 3 byer endnu klarere.

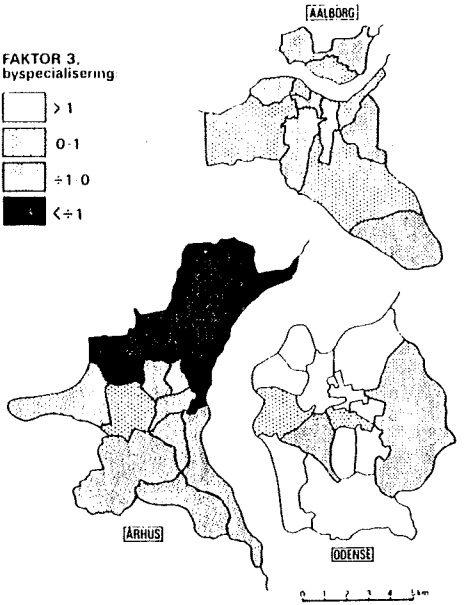


Fig. 6. Kort over faktorværdierne for faktor 3, byspecialisering.
Fig. 6. Map showing the factor values for factor 3, urban specialization.

Ved en sådan drejning af koordinatsystemet bliver fortolkningen af faktor 3's faktorvægte også lettere, og drejningen påvirker ikke nævneværdigt fortolkningen af faktor 2, socio-økonomisk status. Faktor 3' ses derimod nu at skelne mellem universitetsbyen, Århus, og industribyen, Odense. Universitetsbyen Århus har bedre end de andre byer været i stand til at holde på, eller tiltrække, den opvoksende ungdom, de 15–24 årige, og den har som følge af sin størrelse og sit universitet en større beskæftigelse i administration og liberale erhverv, end de to andre. Industribyen Odense har derimod mange arbejdere og relativt mange i den erhvervsaktive aldersgruppe, 25–64 år. Ålborg, der først og fremmest er en oplandsby uden så udpræget specialisering, ligger i midten.

Vi kan således fortolke faktor 3 som byspecialisering. Denne fortolkning af faktor 3 viser, at erhvervsspecialiseringen mellem de tre byer ikke bare er noget påklistret, der skyldes, at Odense har nogle flere industrikvarterer end de andre byer, og at Århus har sit universitetskvarter;

tværtimod gennemsyrer specialiseringen hele byen og påvirker hver eneste zone i de tre byer.

Faktor 3 er først og fremmest en socio-økonomisk faktor, men den har også demografiske træk. Samtidig med at den adskiller de tre byer efter speciale, så har den også lighedspunkter med faktor 3, vækst og mobilitet, i Pedersens (1965 og 1967) analyse af Storkøbenhavn og med Sweetsers (1965) faktor, progeniture, i Helsingfors. I Københavnsanalysen udskilte faktor 3 også de områder, der havde mange unge voksne og få midaldrende og relativt mange funktionærer og beskæftigede i administration og liberale erhverv. I Københavns analyse tolkedes dette som et resultat af, at det især er de yngre aldersgrupper der vandrer og af, at de yngre aldersgrupper i større udstrækning end de ældre er funktionærer. Denne fortolkning kan også holde i denne analyse, idet Århus i årene før 1965 voksede over dobbelt så stærkt som Odense, medens Ålborgs vækstrate lå et sted imellem (se tabel 2).

En sammenligning af de tre byers interne differentiering

Taget over alle 40 zoner i de tre byer har hver af de tre principale faktorer per definition middelværdien 0 og

Tabel 2. Sammenligning mellem Århus, Odense og Ålborgs alders- og erhvervsfordelinger og vækstrate. Byer med forstæder, 1965.

	Århus	Odense	Ålborg
Befolkning 1965	187.000	133.000	123.000
Pct. af befolkningen i aldersgrupperne			
0– 6 år	10,4	10,5	11,1
7–14 år	10,9	11,4	11,8
15–19 år	8,9	8,7	9,1
20–24 år	11,1	8,9	9,2
25–39 år	18,8	18,6	18,5
40–59 år	24,2	25,6	25,3
60–64 år	4,9	5,2	4,9
65+ år	10,8	11,1	10,1
Ialt	100,0	100,0	100,0
0–14 år	21,3	21,9	22,9
15–24 år	20,0	17,6	18,3
25–64 år	47,9	49,4	48,7
65+ år	10,8	11,1	10,1
Pct. af befolkningen ernæret ved			
landbrug	1,0	1,0	1,2
industri og håndværk	30,0	38,6	33,6
byggeindustri	7,4	8,2	9,0
handel og omsætning	14,9	15,3	15,4
transport	8,3	5,7	7,4
administration og liberale erhverv	15,9	12,1	13,1
andet og uoplyst	5,8	6,0	6,6
formue og rente	16,8	13,2	13,7
Ialt	100,0	100,0	100,0
Vækstrate 1960–65 (pct.)	5,9	2,4	3,6

Kilde: Folketællingen 1965.

spredningen 1. Hvis de tre byer var ens, ville dette også være tilfældet med middelværdien og spredningen taget over zonerne i hver by for sig. For at se i hvilket omfang de tre byers indre differentiering afviger fra hinanden har vi i tabel 3 vist middelværdien og spredningen for hver af de tre faktorer og for hver af de tre byer. Forskelle i middelværdi mellem de tre byer er et udtryk for forskellen mellem de tre byers gennemsnitlige statusniveauer, medens forskelle i spredning mellem de tre byer er et udtryk for forskellen mellem byerne i omfanget af den interne differentiering.

For faktor 1, familiestatus, viser tabel 3 at Ålborg har den største og Århus den mindste middelværdi, svarende til at Århus har flest unge husstande, medens Ålborg har færrest. Spredningen af faktor 1 er også mindst i Århus og størst i Ålborg. Årsagen hertil må være at Århus med sit centralt placerede universitet især har flere unge i de centrale bydele, hvor de andre byer har få. Disse forskelle mellem byerne er imidlertid ikke statistisk signifikante, idet ingen af middelværdierne afviger signifikant fra nul, og ingen af spredningerne afviger signifikant fra 1. Spredningen mellem de tre byer er også mindre end den indre differentiering i de enkelte byer. Dette mønster for faktor 1 svarer nøje til den faktor 1 som Janson (1971) fandt for svenske byer.

Middelværdierne for faktor 2, socio-økonomisk status, viser, at Århus i gennemsnit har den højeste status og Ålborg den laveste. Forskellen mellem de tre byer er større end for faktor 1, men ingen af middelværdierne afviger dog signifikant fra nul og den indre differentiering i de tre byer er større end variationen mellem byerne. Også dette mønster svarer til det Janson (1971) fandt for faktor 2 for de svenske byer.

Spredningen af faktor 2 er størst i industribyen Odense og mindst i universitetsbyen Århus. Ingen af de tre spredninger afviger dog signifikant fra 1. Dette er i modstrid med resultaterne for de svenske byer, hvor Janson fandt den største spredning af den socio-økonomiske statusfaktor i universitetsbyerne Uppsala og Lund.

Faktor 3, byspecialisering, er den af de tre faktorer for hvilken forskellen mellem byernes middelværdier er størst. Middelværdierne for både Århus og Odense er signifikant forskellige fra nul, den ene positiv, den anden negativ, medens Ålborgs middelværdi ligger lige midt imellem tæt ved nul, og alle tre middelværdier afviger desuden signifikant fra hinanden to og to.

Spredningen af faktor 3 er for alle tre byer mindre end 1,0, og både for Århus og Ålborg er spredningen signifikant forskellig fra 1,0. Spredningerne for Århus og Ålborg er også signifikant mindre end spredningen for Odense.

Som konklusion kan man sige at industribyen Odense gennemgående har den største indre differentiering, medens universitets- og administrationsbyen Århus har den mindste.

Tabel 3. Middelværdi og spredning for hver af de tre faktorer og for hver af de tre byer.

	Antal zoner	Faktor 1. Familiestatus		Faktor 2. Socio-økonomisk status		Faktor 3. Byspecialisering	
		middel-værdi	spred-ning	middel-værdi	spred-ning	middel-værdi	spred-ning
Århus	14	-0,138	0,795	0,253	0,820	-0,860[1])	0,501[2])
Odense	14	-0,104	0,988	0,061	1,109	0,996[1])	0,820
Ålborg	12	0,282	1,205	0,367	1,029	-0,159	0,476[2])
Ialt	40	0,000	0,981	0,001	0,998	0,001	1,001

[1]) Middelværdien signifikant forskellig fra 0,0 på mere end et 99,9 % niveau. Ingen af de øvrige middelværdier afviger signifikant fra 0,0 på mere end et 88 % niveau (t-test).
[2]) Spredningen signifikant forskellig fra 1,0 på et 99,5 % niveau. Ingen af de øvrige spredninger afviger signifikant fra 1,0 på mere end et 83 % niveau (χ^2-test).

Ved hjælp af Welch-Aspins modificerede t-test har vi testet om middelværdierne for de 3 byer afveg signifikant fra hinanden to og to. Testet viste at dette kun var tilfældet for faktor 3, hvor alle tre middelværdier afveg signifikant fra hinanden på mere end et 95 % niveau.

Ved hjælp af et F-test har vi desuden testet om spredningerne for de 3 byer afveg signifikant fra hinanden. Testet viste at dette kun var tilfældet for faktor 3, for hvilken Odenses spredning afviger signifikant fra Århus' og Ålborgs spredninger på et 95 % niveau. Århus' og Ålborgs spredninger er derimod ikke signifikant forskellige. For faktor 1 er Århus' og Ålborgs spredninger signifikant forskellige på et 90 % niveau.
Kilde for statistiske tabeller: Pearson og Hartley (1962).

Konklusion

Denne analyse har bekræftet, at den struktur, man har fundet i de fleste af den vestlige verdens storbyer, også findes i de største danske provinsbyer. Vi fandt således at byernes befolkningsstruktur kan forklares som samspillet mellem tre faktorer: familiestatus, socio-økonomisk status og byspecialisering, hvor faktor 3, byspecialisering, også ligner den vækst- og vandringsfaktor, der har været beskrevet i en række andre byer.

Det spændende ved faktor 3 er imidlertid at den temmelig præcist adskiller de tre byers zoner fra hinanden, og derved viser at den specialisering der har fundet sted mellem de tre byer (industri i Odense, universitetet i Århus og almindelig oplandshandel i Ålborg) ikke bare er noget påklistret, men påvirker befolkningsstrukturen i hver eneste zone i de tre byer.

Endelig giver analysen, fordi den indeholder zoner fra mere end een by, mulighed for at fremsætte en række hypoteser om hvorledes omfanget af den indre differentiering varierer fra by til by; disse hypoteser kan dog ikke verificeres endeligt på grundlag af denne analyse af kun tre byer:

1. Familiestatusfaktoren følger bystørrelsen, således at den største by der her har haft den største tilvandring også har flest unge husstande og færrest gamle husstande. Den indre differentiering med hensyn til denne faktor er også mindst i den store by og størst i den lille by.

2. De største byer har i gennemsnit den højeste socio-økonomiske status. Den indre differentiering med hensyn til denne faktor følger derimod ikke bystørrelsen, men byspecialiseringen, således at industribyen, Odense, der har en stor andel af arbejdere også har den største indre differentiering, medens det administrative center, Århus, har den mindste indre differentiering.

3. Den indre differentiering med hensyn til faktor 3, byspecialisering, er størst der hvor specialiseringen er kraftigst, og mindst i den almindelige oplandsby.

Alt i alt er konklusionen af vores analyse, at Odense er den mest heterogene og Århus den mest homogene af de tre byer vi har undersøgt.

SUMMARY

This paper analyses the inner differentation in the three largest provincial towns in Denmark by means of a factor analysis of 25 variables characterizing the population and the housing in 40 zones, 14 in Århus (187.000 inh.), 14 in Odense (133.000 inh.) and 12 in Ålborg (123.000 inh.). To be able to compare the inner differentation of the three towns, all 40 zones from the three towns were included in the same factor analysis.

The analysis confirmed that the structure found in most other cities in the western world also is valid for the Danish provincial towns. Thus we found that the population structure of the three towns could be explained as a result of the interplay of three factors: family status, socio-economic status and town specialization, where factor 3, town specialization, also have some similarity to the growth-and-migration-factor found in a number of other towns.

However, the interesting about factor 3 is, that it differentiates the zones of each of the three towns from each other. Thereby it shows that the specialities of the three towns, manufacturing in Odense, university in Århus and ordinary hinterland trade in Ålborg, do not only show up in single zones of the towns, but influence the population structure of every zone in the towns.

Finally the analysis makes it possible to make some hypothesis about how the extent of the inner differentiation differs from town to town; these hypothesis, however, cannot be finally verified on the basis of this investigation of only three towns:

1. The average value of the family status factor is greatest in the largest town, which has experienced the largest immigration, and therefore, also has the highest proportion of young households and the smallest proportion of old households. The inner differentiation with regard to this factor is also smallest in the large town and largest in the small.

2. As an average the largest town has the highest socio-economic status. The inner differentiation with regard to socio-economic status, however, does not follow the town size, but the specialization, so that the manufacturing center, Odense, which has the highest proportion of blue collar workers also has the largest inter-zonal differences in socio-economic status, while the administrative-educational center, Århus, has the smallest inter-zonal differences.

3. The inner differentiation with regard to factor 3, town specialization, is largest in the towns with the most extreme specialization and smallest in the ordinary hinterland town.

Regarding the three towns analysed here, we can conclude that Odense is the most heterogeneous and Århus the most homogeneous town.

LITTERATUR

Berry, Brian J. L. (1964): Cities as Systems within Systems of Cities. Papers and proceedings of the Regional Science Association, 13, 147–163.

Harman, Harry H. (1960): Modern Factor Analysis, Chicago, University of Chicago Press.

Illeris, Sven og Poul Ove Pedersen (1968): Central Places and Functional Regions in Denmark. A Factor Analysis of Telephone Traffic. Geografisk Tidsskrift, 67, 1–18.

Janson, Carl-Gunnar (1971): A Preliminary Report on Swedish Urban Spatial Structure. Economic Geography, 47, 2 (Supplement), 249–257.

Pearson, E. S. and H. D. Hartley (1962): Biometric Tables for Statisticians, 1. Cambridge, University of Cambridge Press.

Pedersen, Poul Ove (1967): Modeller for befolkningsstruktur og befolkningsudvikling i storbyområder – specielt med henblik på Storkøbenhavn. København, Teknisk Forlag.

Pedersen, Poul Ove (1965): An Empirical Model of Population Structure. A Factor Analytic Study of the Population Structure in Copenhagen. Proceedings of first Scandinavian-Polish Regional Science Seminar. Polish Scientific Publishers. Warszawa.

Rees, Philip H. (1970): The Factorial Ecology of Metropolitan Chicago. In B. J. L. Berry and Frank E. Horton (eds.): Geographic Perspectives on Urban Systems. Englewood Cliffs, N.J.

Sweetser, Frank L. (1965a): Factor Structure as Ecological Structure in Helsinki and Boston. Acta Sociologica, 8, 202–25.

Sweetser, Frank L. (1965b): Factorial Ecology. Helsinki, 1960. Demography, 2, 372–86.

Timms, D. W. G. (1971): The Urban Mosaic. Towards a Theory of Residential Differentation. Cambridge, University of Cambridge Press.

Biomedical Computer Programs (1970). Second edition. Health Science Computing Facility, University of California, Los Angeles.

8.3.6 Lidt om maximum likelihood - faktoranalyse

Med fremkomsten af efficiente maksimaliseringsmetoder (f. eks. (Davidon-)Fletcher-Powell's metode) er det blevet muligt at foretage en ML-estimation af faktorvægte. Dette er ud fra en statistisk synsvinkel en mere tilfredsstillende metode end f. eks. principal factor metoden. Endvidere besidder ML-løsningen en skalainvariansegenskab, hvilket også er overordentlig tilfredsstillende.

Vi skal ikke komme ind på de i det væsentlige numerisk-tekniske problemer ved at bestemme ML-løsningen, men mere se på skalainvariansen.

Vi benævner den empiriske kovariansmatrix $\underline{\underline{S}}$ og har under forudsætning af normalitet af observationerne, at $\underline{\underline{S}}$ er Wishart-fordelt med parametre $(n-1, \frac{1}{n-1}\underline{\underline{\Sigma}})$, hvor $\underline{\underline{\Sigma}}$ er lig $D(\underline{X}_i)$, d.v.s. tætheden er

$$c_1 (\det \underline{\underline{S}})^{\frac{1}{2}(n-k-2)} (\det \underline{\underline{\Sigma}})^{-\frac{1}{2}(n-1)} \exp(-\tfrac{1}{2}(n-1)\operatorname{tr}(\underline{\underline{S}}\,\underline{\underline{\Sigma}}^{-1})) \ ,$$

hvor c_1 er en integrationskonstant alene afhængende af n og k. Logaritmen til likelihoodfunktionen er derfor, idet vi udelader de led, der ikke afhænger af $\underline{\underline{\Sigma}}$,

$$\log L(\underline{\underline{\Sigma}}) = -\tfrac{1}{2}(n-1)\log(\det \underline{\underline{\Sigma}}) - \tfrac{1}{2}(n-1)\operatorname{tr}(\underline{\underline{S}}\,\underline{\underline{\Sigma}}^{-1}) \ .$$

Heri introduceres nu den sædvanlige m-faktor model

$$D(\underline{X}) = \underline{\underline{\Sigma}} = \underline{\underline{A}}\,\underline{\underline{A}}' + \underline{\underline{\Delta}} \ ,$$

hvor $\underline{\underline{A}}$ og $\underline{\underline{\Delta}}$ er som i afsnit 8.4. Bemærk i øvrigt, at vi her ikke forudsætter, at $\underline{\underline{\Sigma}}$ har et-taller i diagonalen. Dette giver

$$\log L(\underline{\underline{A}}, \underline{\underline{\Delta}}) = -\tfrac{1}{2}(n-1)\log(\det(\underline{\underline{A}}\,\underline{\underline{A}}' + \underline{\underline{\Delta}}))$$

$$- \tfrac{1}{2}(n-1)\operatorname{tr}(\underline{\underline{S}}(\underline{\underline{A}}\,\underline{\underline{A}}' + \underline{\underline{\Delta}})^{-1}) \ .$$

Maksimalisering af denne funktion med hensyn til $\underline{\underline{A}}$ og $\underline{\underline{\Delta}}$ giver ML-løsningen til vores faktoranalyse. Med hensyn til de tekniske problemer, der er involveret heri henvises til Jöreskog (1967).

Ved partiel differentiation af logaritmen til likelihood-funktionen kommer man efter lange udregninger og algebraiske manipulationer frem til ligningen

$$(*) \qquad \hat{\underline{\underline{A}}} = (\hat{\underline{\underline{\Delta}}} + \hat{\underline{\underline{A}}}\,\hat{\underline{\underline{A}}}')\underline{\underline{S}}^{-1}\,\hat{\underline{\underline{A}}} \ ,$$

se f. eks. Morrison (1967).

Foretager vi en skalatransformation af $\underline{X}$'erne, d.v.s. indfører

$$\underline{Z}_i = \underline{\underline{C}}\,\underline{X}_i \ ,$$

bliver

$$\underline{\underline{S}}_z = \underline{\underline{C}}\,\underline{\underline{S}}_x\,\underline{\underline{C}}' \ ,$$

hvor z og x som fodtegn angiver, om de forskellige størrelser er beregnet på basis af $\underline{Z}_i$ eller $\underline{X}_i$'erne. Med samme notationskonvention fås derfor

$$\hat{\underline{\underline{A}}}_z = (\hat{\underline{\underline{\Delta}}}_z + \hat{\underline{\underline{A}}}_z\hat{\underline{\underline{A}}}_z')\underline{\underline{C}}'^{-1}\underline{\underline{S}}_x^{-1}\underline{\underline{C}}^{-1}\hat{\underline{\underline{A}}}_z \ .$$

Ved præ-multiplikation med $\underline{\underline{C}}^{-1}$ fås

$$(**) \qquad \underline{\underline{C}}^{-1}\hat{\underline{\underline{A}}}_z = [\underline{\underline{C}}^{-1}\hat{\underline{\underline{\Delta}}}_z\underline{\underline{C}}'^{-1} + \underline{\underline{C}}^{-1}\hat{\underline{\underline{A}}}_z(\underline{\underline{C}}^{-1}\hat{\underline{\underline{A}}}_z)']\underline{\underline{S}}_x^{-1}\underline{\underline{C}}^{-1}\hat{\underline{\underline{A}}}_z \ .$$

Ved sammenligning af (*) og (**) ses, at såfremt $\underline{\underline{A}}_x$ er en løsning til (*), da vil

$$\underline{\underline{A}}_z = \underline{\underline{C}}^{-1}\underline{\underline{A}}$$

være en løsning til (**). Med andre ord medfører en skalering af

X'erne (observationerne) med matricen $\underline{\underline{C}}$, at faktorvægtene skaleres med $\underline{\underline{C}}^{-1}$.

Hvis vi opretholder normalitetsforudsætningen, kan vi teste, om faktormodellen holder, d.v.s. teste

$$H_o : \underline{\underline{\Sigma}} = \underline{\underline{\Delta}} + \underline{\underline{A}}\,\underline{\underline{A}}' \quad \text{mod} \quad H_1 : \underline{\underline{\Sigma}} \text{ vilkårlig} .$$

Kvotienttestet vil da være ækvivalent med testet givet ved teststørrelsen

$$Z = (n-1-\tfrac{1}{6}(2k+5)-\tfrac{2}{3}m)\ \log_e \frac{|\hat{\underline{\underline{\Delta}}}+\hat{\underline{\underline{A}}}\,\hat{\underline{\underline{A}}}'|}{|\underline{\underline{S}}|}$$

og forkaste for

$$Z > \chi^2(\tfrac{1}{2}\{(k-m)^2-k-m\}) .$$

Slutteligen skal opmærksomheden henledes på, at der i visse standardprogrammer - f. eks. i BMDP-pakken - er mulighed for at få udført en maximum likelihood faktoranalyse.

Eksempel 8.5 I nedenstående tabel er vist resultatet af dels en principal factor løsning (PCA), dels en maximum likelihood faktoranalyse (ML) og endelig en Little Jiffy løsning (se Kaiser (1970)).

Materialet består af 198 prøver af Portland cement, og hver prøve er analyseret for 15 variable (indhold af forskellige cementmineraler, finhed etc.). De 15 variable er kun anført ved deres respektive numre, da det her ikke er tolkningen, der er essentiel, men alene sammenligningen mellem de tre metoder. I tabellen er vægte, der er numerisk mindre end 0.25 sat lig 0 for at lette overskueligheden.

Vi ser, at de tre metoder giver forbavsende ens resultater. For faktor tre's vedkommende afviger PCA-løsningen noget fra ML- og LJIF-løsningerne.

Variabel	Faktor 1			Faktor 2			Faktor 3		
	PCA	ML	LJIF	PCA	ML	LJIF	PCA	ML	LJIF
1	-0.26	0	0	0.95	0.91	0.95	0	0.36	0
2	0	0	0	-0.98	-1.00	-0.99	0	0	0
3	-0.50	0.93	1.08	0	0	0	-0.40	-0.34	-0.72
4	0.94	-0.78	-0.80	0	0	0	0	-0.62	-0.32
5	0	0.29	0.34	0	0	0	-0.48	0	0
6	0	0	0	0	0	0	0	0	-0.25
7	0	0	0	0	0	0	0	0	0
8	0.53	-0.32	-0.32	0	0	0	0.27	-0.31	0
9	0.90	-0.72	-0.76	0	0	0	0	-0.45	0
10	0	0	0	0	0	0	0.72	0	0
11	0	-0.28	-0.31	0	0	0	0.82	0	0
12	0	0	0	0	0	0	-0.78	0	0
13	-0.73	0	0	0	0	0	0	0.98	0.95
14	-0.86	0.97	1.05	0	0	0	-0.31	0	0
15	0	0.25	0	0.93	0.93	0.92	0	0	-0.35

□

8.3.7 Q-modus analyse

I den form for faktoranalyse, vi hidtil har beskæftiget os med - den såkaldte R-modus analyse - undersøger man korrelationerne mellem de forskellige variable. Individer, prøver etc. regnes for gentagelser, og disse bruges til at estimere de forskellige korrelationer. Kaldes observationerne $\underline{X}_1, \cdots, \underline{X}_n$, og sætter vi

$$\underline{\underline{X}}' = \begin{bmatrix} X_{11} & \cdots & X_{1n} \\ \vdots & & \vdots \\ X_{k1} & \cdots & X_{kn} \end{bmatrix} ,$$

hvor den enkelte række altså svarer til de enkelte variable og de enkelte søjler til individer. Idet vi forudsætter, at målingerne er normerede, så de har middelværdi 0 og variansen 1, fås korrelationsmatricen som

$$\underline{\underline{R}} = \underline{\underline{X}}'\underline{\underline{X}} ,$$

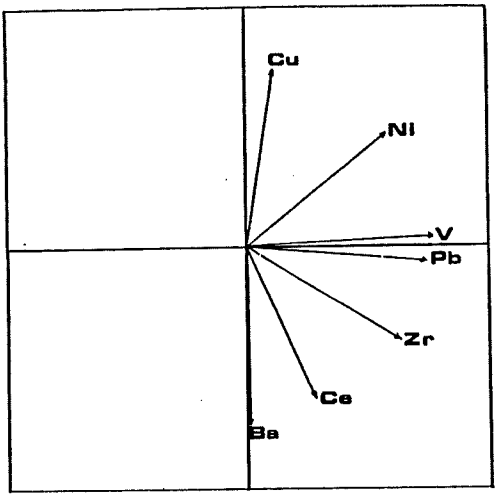
jvf. sætning 2.19. Dualt kan man selvfølgelig definere

$$\underset{=}{Q} = \underset{=}{X}\,\underset{=}{X}' \ ,$$

og opfatte den som et udtryk for korrelationen mellem individer og så udføre en faktoranalyse på disse. Resultatet af en sådan vil blive en klassifikation af individer i grupper af hinanden nærtstående.

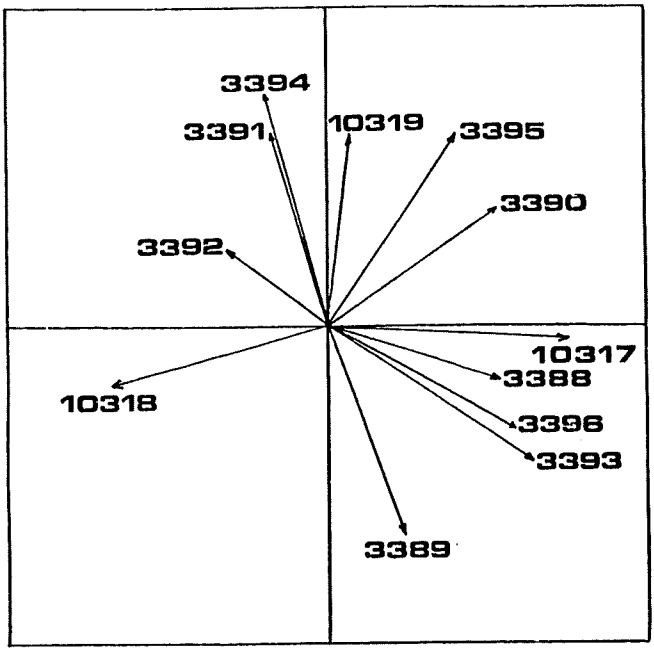
Vi anfører et lille eksempel hentet fra Larsen (1976).

Eksempel 8.6 Vi betragter 12 vaskeprøver indsamlet i Jameson Land i Østgrønland. De er analyseret for 7 elementer, nemlig Cu, Ni, V, Pb, Zr, Ca og Ba. En almindelig R-modus analyse gav, at de to første faktorer beskrev 42% + 37% = 79% af variationen. I nedenstående figur er vist de roterede faktorvægte.



Faktorvægte i R-modus analyse.

Dernæst udførtes en Q-modus analyse som omtalt ovenfor. Dette gav en første faktor, der beskrev 38% af den totale variation, og en anden faktor, der beskrev 26% af den totale variation.



Faktorvægte i Q-modus analyse.

Af figuren med faktorvægtene får vi nu direkte en "sammenligning" af de forskellige prøver. Dette kunne også opnås under R-modus analysen, men da blev man nødsaget til at gå omvejen over factor scores.

Analyser af denne art bruges i mineralprospekteringen ved forsøgene på at få fastlagt, hvilke prøver der må anses for anomale - og dermed interessante.

□

I forbindelse med udførelsen af en Q-modus analyse vil man ofte ende med et meget stort regnearbejde, da Q-matricen er af orden $n \times n$, hvor n er antallet af forsøgspersoner. Man kan da med stor fordel benytte sig af sætningerne i afsnit 1.4.2. Heraf fremgår nemlig, at de fra 0 forskellige egenværdier for R og Q er ens, og der findes en simpel relation mellem egenvektorerne. Da R kun er af orden $k \times k$, og da antallet af variable oftest er væsentligt mindre end antallet af forsøgspersoner, er det muligt at spare en mængde numerisk arbejde.

Til sidst må vi indskyde, at Q-modus analyser ofte ikke foregår på $\underline{\underline{X}}\,\underline{\underline{X}}'$, men på en anden matrix indeholdende nogle mere eller mindre arbitrært valgte similaritetsmål (lighedsmål). Selve teknikken er dog ofte uforandret, og selvfølgelig kan man stadig opnå regnetekniske besparelser ved at anvende den ovenfor omtalte sammenhæng mellem R-modus og Q-modus analyser. For specielle valg af similaritetsmål taler man også ofte om en principal koordinatanalyse.

Et forsøg på en gang at sammenholde begge analyser har man i den såkaldte korrespondanceanalyse, der skyldes franskmanden Benzécri (1973).

8.3.8 Nogle standardprogrammer

En principal komponentanalyse er jo blot en egenværdianalyse af dispersionsmatricen eller en estimeret dispersionsmatrix. En sådan analyse kan derfor laves ved hjælp af et standardprogram til løsning af egenværdiproblemet for en symmetrisk, positivt semidefinit matrix.

Der findes dog også en række standardprogrammer til beregning af principale komponenter. Her kan e.g. nævnes programmerne BMD01M og BMD02M fra BMD-systemet.

BMD01M, PRINCIPAL COMPONENT ANALYSIS, beregner en principal komponentløsning på de standardiserede data, d.v.s. det er den empiriske korrelationsmatrix, der analyseres. Output fra dette program inkluderer korrelationskoefficienter, egenværdier inklusive de kumulerede brøkdele af den totale varians samt egenvektorerne, d.v.s. de principale akser. Endelig anføres en rangordning af hver observation (standardiseret) efter størrelse af de enkelte principale komponenter.

BMD02M, REGRESSION ON PRINCIPAL COMPONENTS, beregner de samme størrelser som BMD01M, og endvidere beregnes regressioner af hver af de afhængige variable på den første, de første to, de første tre og samtlige principale komponenter.

De fleste standardprogrammer til beregning af faktorløsninger bygger på den i denne fremstilling omtalte principale faktorløsning efterfulgt af en rotation.

Et af de mest omfattende systemer er det programkompleks, der er anført i SPSS-manualen (Statistical Package for the Social Sciences). I dette system findes der en række faktoriseringsrutiner. De oftest anvendte er nok principale faktormetoder. Disse findes i to udgaver. En, hvor man blot anvender den almindelige principale faktorløsning, og en, hvor man iterativt estimerer kommunaliteterne ved hjælp af de kvadrerede multiple korrelationskoefficienter, vurderer antallet af nødvendige faktorer, udelukker eventuelt visse, reestimerer kommunaliteterne, etc., indtil forskellen mellem to sæt estimerede kommunaliteter er mindre end en vis grænse.

Blandt en række øvrige findes også en af Rao udviklet mere klassisk statistisk orienteret metode (se Rao (1955)). Her foretages mere sædvanlige estimationer af og test for antallet af nødvendige faktorer m.v.

Af ortogonale rotationsprincipper findes tre, nemlig quartimax, varimax (se p. 8.28) og equimax. Endvidere findes en procedure, der udfører en såkaldt oblique rotation (efter oblimin-princippet).

Beregning af faktorværdier foregår efter et princip, der er beslægtet med det, der er omtalt i afsnit 8.3.7.

Også BMD-programmet BMDO8M, FACTOR ANALYSIS, er ganske omfattende. Faktoriseringsrutinerne er dog alle af principal faktortypen. De opererer på såvel korrelations- som dispersionsmatricer. Der er muligheder for forskellige former for kommunalitetsestimater, og den ovenfor omtalte iterative estimationsprocedure kan bruges.

Der findes en række rotationsprincipper, såvel ortogonale (bl.a. quartimax og varimax) som "oblique" (oblimin-typer).

Beregning af faktorværdier foregår efter samme principper som omtalt i afsnit 8.3.7.

I BMDP-pakkens faktoranalyseprogram kan man som nævnt også få udført en ML-estimation.

SSP-sampleprogrammet FACTO laver en principal faktorløsning og roterer faktorerne ved varimax-metoden. Programmet er i det store og hele identisk med det gamle faktoranalyseprogram fra BMD-systemet, nemlig BMDO3M. Output omfatter de sædvanlige størrelser, dog ikke faktorværdier (factor scores). Nogle anvendelser skal gennemgås nedenfor.

Programmet FACTO kalder en brugerrutine DATA og 5 rutiner fra SSP-pakken, der alle p.t. er lagt ind under WATFIV compileren. Dette muliggør en ret hurtig afvikling af et program.

P. 8.57-58 er anført en programudskrift for FACTO og DATA. I den anførte version udføres en faktoranalyse for op til 35 variable og op til 99.999 observationer. I øvrigt henvises til p. 429 i SSP-manualen.

Der kræves blot et enkelt styrekort, der udfyldes som angivet øverst p. 8.59.

```
C        SAMPLE MAIN PROGRAM FOR FACTOR ANALYSIS - FACTO                        01900000
C                                                                               01910000
C        PURPOSE                                                                01920000
C           (1) READ THE PROBLEM PARAMETER CARD, (2) CALL FIVE SUBROU-          01930000
C           TINES TO PERFORM A PRINCIPAL COMPONENT SOLUTION AND THE             01940000
C           VARIMAX ROTATION OF A FACTOR MATRIX, AND (3) PRINT THE              01950000
C           RESULTS.                                                            01960000
C                                                                               01970000
C        REMARKS                                                                01980000
C           NONE                                                                01990000
C                                                                               02000000
C        SUBROUTINES AND FUNCTION SUBPROGRAMS REQUIRED                          02010000
C           CORRE  (WHICH, IN TURN, CALLS THE SUBROUTINE NAMED DATA.)           02020000
C           EIGEN                                                               02030000
C           TRACE                                                               02040000
C           LOAD                                                                02050000
C           VARMX                                                               02060000
C                                                                               02070000
C        METHOD                                                                 02080000
C           REFER TO 'BMD COMPUTER PROGRAMS MANUAL', EDITED BY W. J.            02090000
C           DIXON, UCLA, 1964.                                                  02100000
C                                                                               02110000
C     ******************************************************************        02120000
C                                                                               02130000
C     THE FOLLOWING DIMENSIONS MUST BE GREATER THAN OR EQUAL TO THE             02140000
C     NUMBER OF VARIABLES, M..                                                  02150000
C                                                                               02160000
      1     DIMENSION B(35),D(35),S(35),T(35),XBAR(35)                          02170000
      2     DIMENSION X(1)                                                      
C                                                                               02180000
C     THE FOLLOWING DIMENSION MUST BE GREATER THAN OR EQUAL TO THE              02190000
C     PRODUCT OF M*M..                                                          02200000
C                                                                               02210000
      3     DIMENSION V(1225)                                                   02220000
C                                                                               02230000
C     THE FOLLOWING DIMENSION MUST BE GREATER THAN OR EQUAL TO                  02240000
C     (M+1)*M/2..                                                               02250000
C                                                                               02260000
      4     DIMENSION R(630)                                                    02270000
C                                                                               02280000
C     THE FOLLOWING DIMENSION MUST BE GREATER THAN OR EQUAL TO 51..             02290000
C                                                                               02300000
      5     DIMENSION TV(51)                                                    02310000
C                                                                               02320000
C     ******************************************************************        02330000
C                                                                               02340000
C        IF A DOUBLE PRECISION VERSION OF THIS ROUTINE IS DESIRED, THE          02350000
C        C IN COLUMN 1 SHOULD BE REMOVED FROM THE DOUBLE PRECISION              02360000
C        STATEMENT WHICH FOLLOWS.                                               02370000
C                                                                               02380000
C     DOUBLE PRECISION XBAR,S,V,R,D,B,T,TV                                      02390000
C                                                                               02400000
C        THE C MUST ALSO BE REMOVED FROM DOUBLE PRECISION STATEMENTS            02410000
C        APPEARING IN OTHER ROUTINES USED IN CONJUNCTION WITH THIS              02420000
C        ROUTINE.                                                               02430000
C                                                                               02440000
C        ******************************************************************     02450000
C                                                                               02460000
      6       1 FORMAT(21H1FACTOR ANALYSIS.....,A4,A2//3X,12HNO. OF CASES,4X,I6/3X,02470000
             1 16HNO. OF VARIABLES,I6/)                                         02480000
      7       2 FORMAT(6H0MEANS/(8F15.5))                                       02490000
      8       3 FORMAT(20H0STANDARD DEVIATIONS/(8F15.5))                        02500000
      9       4 FORMAT(25H0CORRELATION COEFFICIENTS)                            02510000
     10       5 FORMAT(4H0ROWI3/(10F12.5))                                      02520000
     11       6 FORMAT(1H0/12H EIGENVALUES/(10F12.5))                           02530000
     12       7 FORMAT(37H0CUMULATIVE PERCENTAGE OF EIGENVALUES/(10F12.5))      02540000
     13       8 FORMAT(1H0/13H EIGENVECTORS)                                    02550000
     14       9 FORMAT(7H0VECTOI3/(10F12.5))                                    02560000
     15      10 FORMAT(1H0/16H FACTOR MATRIX (,I3,9H FACTORS))                  02570000
     16      11 FORMAT(9H0VARIABLEI3/(10F12.5))                                 02580000
     17      12 FORMAT(1H0/10H ITERATION,7X,9HVARIANCES/8H   CYCLE)             02590000
     18      13 FORMAT(I6,F20.6)                                               02600000
     19      14 FORMAT(1H0/24H ROTATED FACTOR MATRIX (I3,9H FACTORS))          02610000
     20      15 FORMAT(9H0VARIABLEI3/(10F12.5))                                02620000
     21      16 FORMAT(1H0/23H CHECK ON COMMUNALITIES//9H VARIABLE,7X,8HORIGINAL,02630000
             1 12X,5HFINAL,10X,10HDIFFERENCE)                                   02640000
     22      17 FORMAT(I6,3F18.5)                                              02650000
     23      18 FORMAT(A4,A2,I5,I2,F6.0)                                       02660000
     24      19 FORMAT(5H0ONLY,I2,30H FACTOR RETAINED.  NO ROTATION)           02670000
C                                                                               02680000
C     ******************************************************************        02690000
C                                                                               02700000
C        READ PROBLEM PARAMETER CARD                                            02710000
C                                                                               02720000
     25   100 READ (5,18) PR,PR1,N,M,CON                                        02730000
C           PR.........PROBLEM NUMBER (MAY BE ALPHAMERIC)                       02740000
C           PR1........PROBLEM NUMBER (CONTINUED)                               02750000
C           N..........NUMBER OF CASES                                          02760000
C           M..........NUMBER OF VARIABLES                                      02770000
C           CON........CONSTANT USED TO DECIDE HOW MANY EIGENVALUES             02780000
C                      TO RETAIN                                                02790000
C                                                                               02800000
     26       WRITE (6,1) PR,PR1,N,M                                            02810000
C                                                                               02820000
     27       IO=0                                                              02830000
     28       X(1)=0.0                                                          
C                                                                               02850000
     29       CALL CORRE (N,M,IO,X,XBAR,S,V,R,D,B,T)                            02860000
C                                                                               02870000
C        PRINT MEANS                                                            02880000
C                                                                               02890000
     30       WRITE (6,2) (XBAR(J),J=1,M)                                       02900000
C                                                                               02910000
C        PRINT STANDARD DEVIATIONS                                              02920000
```

```
      C                                                                 02930000
   31       WRITE (6,3) (S(J),J=1,M)                                    02940000
      C                                                                 02950000
      C     PRINT CORRELATION COEFFICIENTS                              02960000
      C                                                                 02970000
   32       WRITE (6,4)                                                 02980000
   33       DO 120 I=1,M                                                02990000
   34       DO 110 J=1,M                                                03000000
   35       IF(I-J) 102, 104, 104                                       03010000
   36   102 L=I+(J*J-J)/2                                               03020000
   37       GO TO 110                                                   03030000
   38   104 L=J+(I*I-I)/2                                               03040000
   39   110 D(J)=R(L)                                                   03050000
   40   120 WRITE (6,5) I,(D(J),J=1,M)                                  03060000
      C                                                                 03070000
   41       MV=0                                                        03080000
   42       CALL EIGEN (R,V,M,MV)                                       03090000
      C                                                                 03100000
   43       CALL TRACE (M,R,CON,K,D)                                    03110000
      C                                                                 03120000
      C     PRINT EIGENVALUES                                           03130000
      C                                                                 03140000
   44       DO 130 I=1,K                                                03150000
   45       L=I+(I*I-I)/2                                               03160000
   46   130 S(I)=R(L)                                                   03170000
   47       WRITE (6,6) (S(J),J=1,K)                                    03180000
      C                                                                 03190000
      C     PRINT CUMULATIVE PERCENTAGE OF EIGENVALUES                  03200000
      C                                                                 03210000
   48       WRITE (6,7) (D(J),J=1,K)                                    03220000
      C                                                                 03230000
      C     PRINT EIGENVECTORS                                          03240000
      C                                                                 03250000
   49       WRITE (6,8)                                                 03260000
   50       L=0                                                         03270000
   51       DO 150 J=1,K                                                03280000
   52       DO 140 I=1,M                                                03290000
   53       L=L+1                                                       03300000
   54   140 D(I)=V(L)                                                   03310000
   55   150 WRITE (6,9) J,(D(I),I=1,M)                                  03320000
      C                                                                 03330000
   56       CALL LOAD (M,K,R,V)                                         03340000
      C                                                                 03350000
      C     PRINT FACTOR MATRIX                                         03360000
      C                                                                 03370000
   57       WRITE (6,10) K                                              03380000
   58       DO 180 I=1,M                                                03390000
   59       DO 170 J=1,K                                                03400000
   60       L=M*(J-1)+I                                                 03410000
   61   170 D(J)=V(L)                                                   03420000
   62   180 WRITE (6,11) I,(D(J),J=1,K)                                 03430000
      C                                                                 03440000
   63       IF(K-1) 185, 185, 188                                       03450000
   64   185 WRITE (6,19) K                                              03460000
   65       GO TO 100                                                   03470000
      C                                                                 03480000
   66   188 CALL VARMX(M,K,V,NC,TV,B,T,D,IJR)                           03500000
      C                                                                 03510000
      C     PRINT VARIANCES                                             03520000
      C                                                                 03530000
   67       NV=NC+1                                                     03540000
   68       WRITE (6,12)                                                03550000
   69       DO 190 I=1,NV                                               03560000
   70       NC=I-1                                                      03570000
   71   190 WRITE (6,13) NC,TV(I)                                       03580000
      C                                                                 03590000
      C     PRINT ROTATED FACTOR MATRIX                                 03600000
      C                                                                 03610000
   72       WRITE (6,14) K                                              03620000
   73       DO 220 I=1,M                                                03630000
   74       DO 210 J=1,K                                                03640000
   75       L=M*(J-1)+I                                                 03650000
   76   210 S(J)=V(L)                                                   03660000
   77   220 WRITE (6,15) I,(S(J),J=1,K)                                 03670000
      C                                                                 03680000
      C     PRINT COMMUNALITIES                                         03690000
      C                                                                 03700000
   78       WRITE (6,16)                                                03710000
   79       DO 230 I=1,M                                                03720000
   80   230 WRITE (6,17) I,B(I),T(I),D(I)                               03730000
   81       GO TO 100                                                   03740000
   82       END

   83       SUBROUTINE DATA(M,D)
   84       DIMENSION D(1)
   85       CALL WATCON
   86     1 FORMAT(F6.0,6F7.0)
   87       READ(5,1)(D(I),I=1,M)
   88       RETURN
   89       END
```

Kolonne	Indhold
1-6	Navn på analyse (gerne alfamerisk).
7-11	Antal observationer.
12-13	Antal variable.
14-19	Grænse for størrelse af egenværdier i korrelationsmatricen. Kun egenværdier større end eller lig denne grænse medtages i analysen. Der skal specificeres et decimalpunktum.

Data indlæses efter FORMAT-statement 1 i subrutinen DATA, d.v.s. i den anførte version efter

```
1     FORMAT (F6.0, 6F7.0) .
```

Vi viser nu et eksempel på kørsel med FACTO.

Eksempel 8.7 Vi betragter de data, der er anført i eksempel 8.1 og vil anvende FACTO ved en faktoranalyse af dem.

Input er anført nedenfor. Kørslen er benævnt "KASSE". De øvrige angivelser på styrekortet er åbenbare.

```
                    Kolonne
                    1     7    12 14
                    ↓     ↓    ↓  ↓
Styrekort →         KASSE 0002507 0001.0
                    3.760  3.660  0.540  5.275  9.768 13.741  4.782 ← Datakort 1
                    8.590  4.990  1.340 10.022  7.500 10.162  2.130
                    6.220  6.140  4.520  9.842  2.175  2.732  1.089
                    7.570  7.280  7.070 12.662  1.791  2.101  0.822
                    9.030  7.080  2.590 11.762  4.539  6.217  1.276
                    5.510  3.980  1.300  6.924  5.326  7.304  2.403
                    3.270  0.620  0.440  3.357  7.629  8.838  3.389
                    8.740  7.000  3.310 11.675  3.529  4.757  1.119
                    9.640  9.490  1.030 13.567 13.133 18.519  2.354
                    9.730  1.330  1.000  9.871  9.871 11.064  3.704
                    8.590  2.980  1.170  9.170  7.851  9.909  2.616
                    7.120  5.490  3.680  9.716  2.642  3.430  1.189
                    4.690  3.010  2.170  5.983  2.760  3.554  2.013
                    5.510  1.340  1.270  5.808  4.566  5.382  3.427
                    1.660  1.610  1.570  2.799  1.783  2.087  3.716
                    5.900  5.760  1.550  8.388  5.395  7.497  1.973
                    9.840  9.270  1.510 13.604  9.017 12.668  1.745
                    8.390  4.920  2.540 10.053  3.956  5.237  1.432
                    4.940  4.380  1.030  6.678  6.494  9.059  2.807
                    7.230  2.30   1.770  7.790  4.393  5.374  2.274
                    9.460  7.310  1.040 11.999 11.579 16.182  2.415
                    9.550  5.350  4.250 11.742  2.766  3.509  1.054
                    4.940  4.520  4.500  8.067  1.793  2.103  1.292
                    8.210  3.080  2.420  9.097  3.753  4.657  1.719
                    9.410  6.440  5.110 12.495  2.446  3.103  0.914 ← Datakort 25
```

Input til eksemplet "KASSE".

Output er anført på p. 8.61-62.

Vektoren af middelværdier og standardafvigelser kræver ingen forklaring, og de svarer fuldstændigt til de i eksempel 8.1 (p. 8.11) viste. Tilsvarende er korrelationsmatricen lig den i eksempel 8.3 anførte.

Da vi i styrekortet har specificeret, at vi kun ønsker at medtage egenvektorer svarende til egenværdier større end 1, får vi kun de to største egenværdier og tilhørende egenvektorer ud, jvf. eksempel 8.3, p. 8.30.

Den estimerede principale faktorløsning svarer til den i eksempel 8.3 anførte, blot er orienteringen af den anden faktor den modsatte.

Tilsvarende ser vi, at den varimax-roterede faktorløsning minder meget om den roterede løsning, vi kom frem til ad "visuel" vej p. 8.32, stadig bortset fra orienteringen af anden faktor.

Det fremgår i øvrigt, at der er brugt 5 iterationer for at nå frem til den endelige varimax-løsning. De størrelser, der er benævnt "variances", svarer til udtrykket p. 8.28, som er den størrelse, der skal maksimaliseres ved varimax-metoden.

Endelig er der anført et såkaldt "check on communalities". Dette består, som også nævnt i eksempel 8.4, blot i en beregning af kommunaliteterne på den oprindelige faktorløsning og på den roterede. Disse skal være ens.

□

FACTOR ANALYSIS.....KASSE

NO. OF CASES 25
NO. OF VARIABLES 7

MEANS

7.09999 4.77319 2.34879 9.13383 5.45819 7.16743 2.34615

STANDARD DEVIATIONS

2.32380 2.41778 1.66556 3.01782 3.27326 4.55808 1.61050

CORRELATION COEFFICIENTS

ROW 1
1.00000 0.58026 0.20113 0.91126 0.28333 0.28655 -0.53320

ROW 2
0.58026 1.00000 0.36379 0.83375 0.16583 0.26107 -0.60872

ROW 3
0.20113 0.36379 1.00000 0.43857 -0.70418 -0.68054 -0.64884

ROW 4
0.91126 0.83375 0.43857 1.00000 0.16304 0.20229 -0.67554

ROW 5
0.28333 0.16583 -0.70418 0.16304 1.00000 0.99021 0.42721

ROW 6
0.28655 0.26107 -0.68054 0.20229 0.99021 1.00000 0.35713

ROW 7
-0.53320 -0.60872 -0.64884 -0.67554 0.42721 0.35713 1.00000

EIGENVALUES

3.39462 2.80546

CUMULATIVE PERCENTAGE OF EIGENVALUES

0.48494 0.88572

EIGENVECTORS

VECTOR 1
0.40529 0.43158 0.38544 0.49389 -0.12771 -0.09680 -0.48094

VECTOR 2
-0.29290 -0.22244 0.35588 -0.23227 -0.57510 -0.58000 -0.13030

FACTOR MATRIX (2 FACTORS)

VARIABLE 1
0.74673 -0.49059

VARIABLE 2
0.79516 -0.37258

VARIABLE 3
0.71015 0.59608

VARIABLE 4
0.90996 -0.38904

VARIABLE 5
-0.23530 -0.96327

VARIABLE 6
-0.17835 -0.97148

VARIABLE 7
-0.88610 -0.21824

Output

ITERATION CYCLE	VARIANCES
0	0.241945
1	0.351977
2	0.351977
3	0.351977
4	0.351977
5	0.351977

ROTATED FACTOR MATRIX (2 FACTORS)

VARIABLE 1	
0.87886	-0.16090
VARIABLE 2	
0.87749	-0.03335
VARIABLE 3	
0.42181	0.82565
VARIABLE 4	
0.98963	-0.00379
VARIABLE 5	
0.15858	-0.97883
VARIABLE 6	
0.21422	-0.96420
VARIABLE 7	
-0.73106	-0.54622

CHECK ON COMMUNALITIES

VARIABLE	ORIGINAL	FINAL	DIFFERENCE
1	0.79828	0.79828	0.00000
2	0.77110	0.77110	0.00000
3	0.85963	0.85963	0.00000
4	0.97939	0.97939	0.00000
5	0.98325	0.98325	0.00000
6	0.97558	0.97558	0.00000
7	0.83281	0.83281	0.00000

Referencer til kapitel 8

Agterberg, F.P.: Geomathematics. Mathematical background and geo-science applications. Elsevier, Amsterdam 1973.

Anderson, T.W.: An introduction to multivariate statistical analysis. John Wiley, New York 1958.

Benzécri, J.P.: L'Analyse des Données. 2, L'Analyse des Correspondances. Dunod, Paris 1973.

Cattell, Raymond: Factor analysis: An introduction to essentials. I. The purpose and underlying models. II. The role of factor analysis in research. Biometrics 21 (1965), pp. 190-215 & 405-435.

Davis, John C.: Statistics and data analysis in geology. John Wiley, New York 1973.

Dixon, J.W. (ed.): Biomedical Computer Programs. University of California Press, Los Angeles 1973.

Dixon, J.W. (ed.): BMDP. Biomedical Computer Programs. University of California Press, Los Angeles 1975.

Dwyer, Paul S.: The contribution of an orthogonal multiple facto solution to multiple correlation. Psyck. 4 (1939), pp. 163-171.

Harman, Harry H.: Modern Factor Analysis (sec. ed.). The University of Chicago Press. Chicago 1967.

Hotelling, Harold: Analysis of a complex of statistical variables into principal components. J. Educ. Psych. 24 (1933), pp. 417-441 & 498-520.

Jöreskog, K.G.: Some Contributions to Maximum Likelihood Factor Analysis. Psychometrika, 32, 1967.

Kaiser, H.F.: The varimax criterion for analytic rotation in factor analysis. Psych. 23 (1958), pp. 187-200.

Larsen, P.M.: Geokemisk oversigtsprospektering. Multivariable Statistiske Metoders anvendelighed ved interpretation af regionale geokemiske data. Examensprojekt, IMSOR 1976.

Lawley, D.N.: The estimation of factor loadings by the method of maximum likelihood. Proc. Roy. Soc. Edin. 60 (1940), pp. 64-82.

Morrison, Donald F.: Multivariate Statistical Methods. McGraw-Hill, New York 1967.

Nie, N.H. et al.: Statistical Package for the Social Sciences. McGraw-Hill, New York 1970.

Rao, C.R.: Estimation and tests of significance in factor analysis. Psych. 20 (1955), pp. 93-111.

Spearman, C.: General intelligence objectively determined and measured. Am. Journ. of Psych. 15 (1904) pp. 201-293.

System/360 Scientific Subroutine Package. Version III (fifth ed.). International Business Machines Corporation 1970.

Thurstone, L.L.: Multiple factor analysis. Psych. Rev. 38 (1931), pp. 406-427.

KAPITEL 9

Statistisk analyse af tidsrækker

I dette kapitel skal vi kun give en lille indføring i den statistiske analyse af tidsrækker. Vi vil indledningsvis gennemgå - uden hensyntagen til læserens eventuelle ønsker om at opretholde blot et minimum af matematisk stringens - udvalgte dele af teorien om Fourier-transformationen og i et senere afsnit give en kort gennemgang - eller rettere resumé - af den nødtørftigste teori for stokastiske processer.

Hvad angår selve tidsrækkeanalysen, falder kapitlet i 5 afsnit: et om den såkaldt klassiske analyse, et om en-dimensional spektralanalyse, et om filtrering, et om fler-dimensional spektralanalyse og endelig et om Box-Jenkins' metode.

Som nævnt er det ikke hensigten med dette kapitel at tilstræbe en grundig og systematisk gennemgang af de centrale dele af teorien. Formålet har snarere været at give læseren dels et indtryk af de metoder, der oftest anvendes ved løsningen af problemer, hvor de i en tidsmæssig sekvens indkomne data ikke nødvendigvis er stokastisk uafhængige, og dels at gøre ham så fortrolig med den gængse notation, at det ikke vil volde de store vanskeligheder at frekventere den existerende litteratur.

Dette kræver til gengæld, at den gængse notation følges i kapitlet, og det har medført visse uundgåelige inkonsistenser i valget af symboler.

9.1 Fourier-transformationen, forskydningsoperatorer og lineære systemer

I dette indledende afsnit vil vi give en nødtørftig gennemgang af det matematiske begrebsapparat, som er nødvendigt for at give en nogenlunde rimelig fremstilling af teorien for tidsrækker.

9.1.1 Fourier-transformationen

Fourier-transformationen S(f) til et signal s(t) kan opfattes som generalisation af Fourier-rækkefremstillingen af en periodisk funktion.

Uden at komme ind på hele problematikken vedrørende eksistensen af integralerne definerer vi den Fourier-transformerede til s ved

$$S(f) = \int_{-\infty}^{\infty} s(t)\ e^{-i2\pi ft}\ dt \qquad (1)$$

Det skal bemærkes, at man i litteraturen møder en række forskellige versioner af definitionen. Divergenserne vedrører koefficienten 2π til leddet -ift. Den udelades af en række forfattere.

Signalet s kan bestemmes ved hjælp af følgende inversionsformel

$$s(t) = \int_{-\infty}^{\infty} S(f)\ e^{i2\pi ft}\ df \qquad (2)$$

Da

$$e^{i2\pi ft} = \cos(2\pi ft) + i\ \sin(2\pi ft)\ , \qquad (3)$$

ser vi af (2), at den Fourier-transformerede S angiver koefficienten til den harmoniske svingning med frekvens f i en opspaltning af signalet efter frekvenser.

Betragtes et signal $s(t)$, der kun er forskelligt fra 0 for værdierne

$$\cdots, -n\Delta, \cdots, -\Delta, 0, \Delta, \cdots, n\Delta, \cdots,$$

og sættes $s(k\Delta) = s_k$, defineres helt analogt den diskrete Fourier-transformation ved

$$S(f) = \Delta \sum_{k=-\infty}^{\infty} s_k \, e^{-i2\pi kf\Delta} \quad , \quad -\frac{1}{2\Delta} \leq f \leq \frac{1}{2\Delta} \, .$$

Inversionsformlen bliver

$$s_k = \int_{-\frac{1}{2\Delta}}^{\frac{1}{2\Delta}} S(f) \, e^{i2\pi kf\Delta} df \, .$$

Vi skal ikke komme noget videre ind på problemerne for den diskrete transformation; men blot stedse betone, at der gælder analoge resultater for denne. Dette kan i nogen grad begrundes med den sammenhæng, der eksisterer mellem transformationen af et "samplet" signal fra en kontinuert funktion, jvf. p. 9.29.

Vi giver nu en række eksempler, som skal belyse begrebet.

Eksempel 9.1 Hvis signalet er

$$s(t) = e^{-|t|}$$

bliver

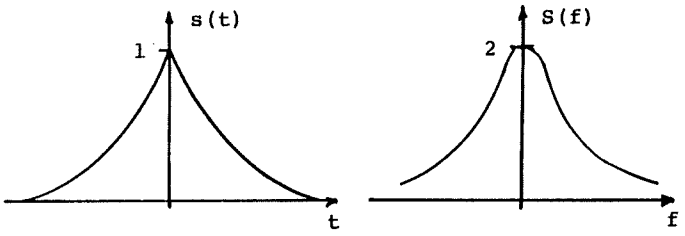
$$S(f) = \int_{-\infty}^{\infty} e^{-|t|}\, e^{-i2\pi ft}\, dt$$

$$= \int_{0}^{\infty} e^{-(1+i2\pi f)t}\, dt + \int_{-\infty}^{0} e^{(1-i2\pi f)t}\, dt$$

$$= \frac{1}{1+i2\pi f} + \frac{1}{1-i2\pi f}$$

$$= \frac{2}{1+(2\pi f)^2}$$

Graferne for s og S er



□

I det ovenstående tilfælde så vi, at S var reel. At dette ikke altid er tilfældet, fremgår af

Eksempel 9.2 Med

$$s(t) = e^{-t}\, I_{[0,\infty[}(t)\ ,$$

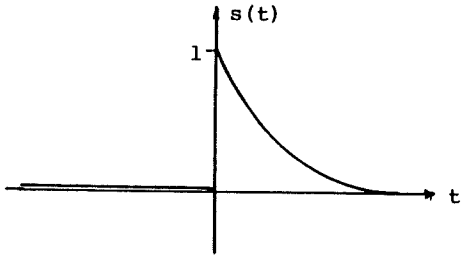
bliver

$$S(f) = \int_{0}^{\infty} e^{-t}\, e^{-i2\pi ft}\, dt$$

$$= \frac{1}{1+i2\pi f}$$

$$= \frac{1}{1+(2\pi f)^2} - i\,\frac{2f}{1+(2\pi f)^2}$$

Grafen for s er



□

Ovenstående to eksempler indicerer, at vi har følgende

Sætning 9.1 Hvis signalet s er symmetrisk, er den Fourier-transformerede S givet ved (1) reel.

Bevis Ses let ved anvendelse af definitionen og Eulers formel (3).

□

Eksempel 9.3 For

$$s(t) = \begin{cases} 1 - \frac{|t|}{T} & , \quad |t| \le T \\ 0 & , \quad |t| > T \end{cases},$$

som vi senere vil benævne et Bartlett-vindue, fås

$$S(f) = \int_{-T}^{T} (1 - \frac{|t|}{T})\, e^{-i2\pi ft}\, dt$$

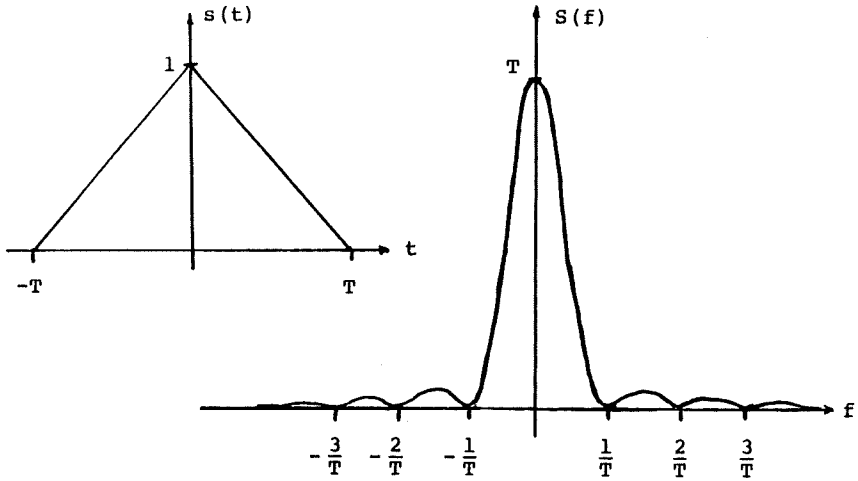
$$= 2\int_{0}^{T} (1 - \frac{|t|}{T})\, \cos(2\pi ft)\, dt$$

$$= \frac{1}{2\pi^2 Tf^2}\,[1 - \cos(2\pi fT)]$$

dvs.

$$S(f) = T\left[\frac{\sin(\pi Tf)}{\pi Tf}\right]^2 , \quad -\infty \le f \le \infty .$$

Graferne er anført nedenfor.



□

Eksempel 9.4 For det såkaldte Hanning-vindue

$$s(t) = \begin{cases} \frac{1}{2} + \frac{1}{2}\cos\frac{\pi t}{T} & , \quad |t| \le T \\ 0 & , \quad |t| > T \end{cases} ,$$

fås

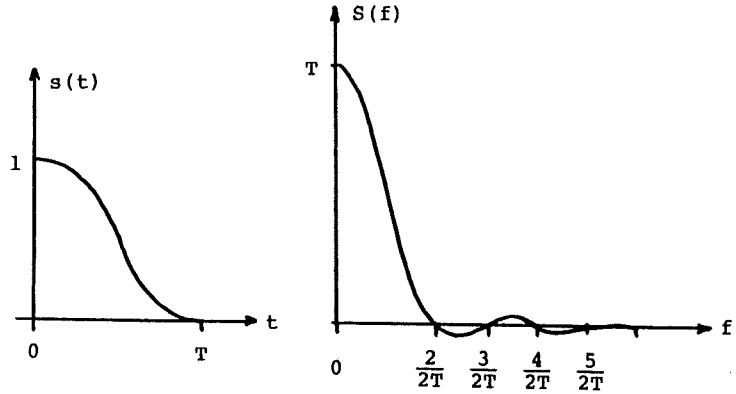
$$S(f) = 2\int_0^T \left(\frac{1}{2} + \frac{1}{2}\cos\frac{\pi t}{T}\right)\cos(2\pi ft)\,dt$$

$$= \frac{1}{2\pi f}\sin(2\pi fT) + \tfrac{1}{2}\int_0^T \left\{\cos\left(\frac{\pi t}{T} + 2\pi ft\right) + \cos\left(\frac{\pi t}{T} - 2\pi ft\right)\right\} dt$$

dvs.

$$S(f) = T\,\frac{\sin(2\pi fT)}{2\pi fT}\,\frac{1}{1-4f^2T^2} , \quad -\infty \le f \le \infty .$$

Graferne er



NB: Begge grafer er symmetriske om anden aksen.

□

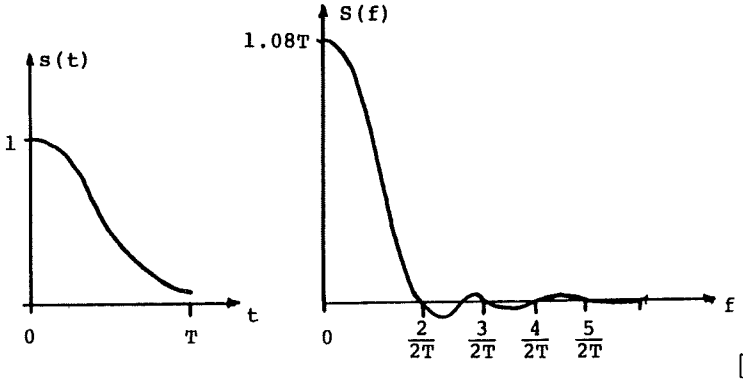
Eksempel 9.5 For Hamming-vinduet

$$s(t) = \begin{cases} 0.54 + 0.46 \cos \frac{\pi t}{T} & , \quad |t| \leq T \\ 0 & , \quad |t| > T \end{cases} ,$$

fås

$$S(f) = T \frac{\sin(2\pi fT)}{2\pi fT} \cdot \frac{1.08 - 0.64\, f^2T^2}{1 - 4f^2T^2} \quad , \quad -\infty \leq f \leq \infty \, .$$

Graferne er symmetriske og vist nedenfor.



□

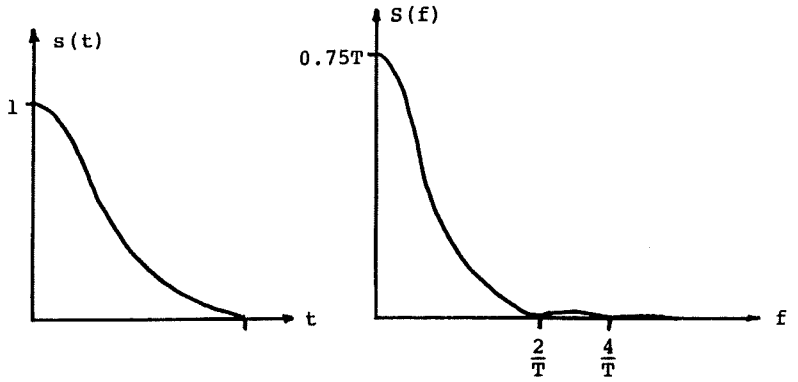
Eksempel 9.6 For

$$s(t) = \begin{cases} 1-6\left(\frac{t}{T}\right)^2 + 6\left(\frac{|t|}{T}\right)^3 , & |t| \le \frac{T}{2} \\ 2\left(1 - \frac{|t|}{T}\right)^3 , & \frac{T}{2} < |t| \le T \\ 0 , & |t| > T \end{cases} ,$$

der også kaldes Parzen-vinduet, fås

$$S(f) = \frac{3}{4} T\left(\frac{\sin(\pi fT/2)}{\pi fT/2}\right)^4 , \quad -\infty \le f \le \infty .$$

Graferne er symmetriske og vist nedenfor.



□

Eksempel 9.7 Vi betragter nu det rektangulære vindue

$$s(t) = a\, I_{[-T,T]}(t) .$$

Fourier integralet bliver

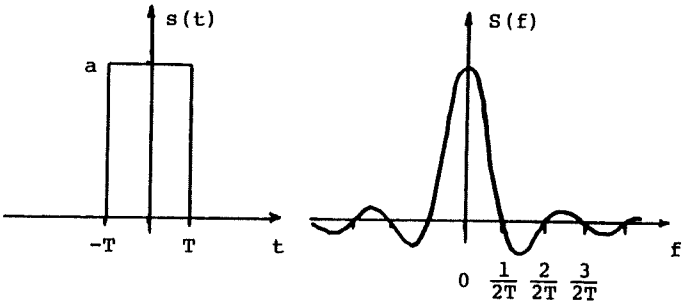
$$S(f) = a\int_{-T}^{T} e^{-i2\pi ft}\, dt$$

$$= a\int_{-T}^{T} [\cos 2\pi ft - i \sin 2\pi ft]\, dt$$

dvs.

$$S(f) = a\,\frac{2 \sin 2\pi fT}{2\pi f} .$$

Graferne er



□

Inden vi fortsætter, må vi indføre en ny "funktion" - eller rettere generaliseret funktion eller distribution - nemlig Dirac's δ-funktion. Den defineres ved

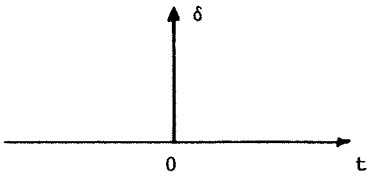
$$\int_{-\infty}^{\infty} \delta(t)\, f(t)\, dt = f(0) \qquad (4)$$

$$\int_{-\infty}^{\infty} \delta(t)\, dt = 1 \qquad (5)$$

$$\delta(t) = 0 \quad \text{for} \quad t \neq 0 \qquad (6)$$

Her følger (5) klart af (4). δ-funktionen er som nævnt ikke en funktion i almindelig forstand. Af (5) og (6) ses, at man kan sige, at δ er 0 overalt undtagen i 0, hvor den er så "stor", at integralet bliver 1. En nærmere og mere tilfredsstillende omtale kan findes i Lighthill (1958).

"Grafen" for δ afbildes sædvanligt som en spids af længden 1 i 0



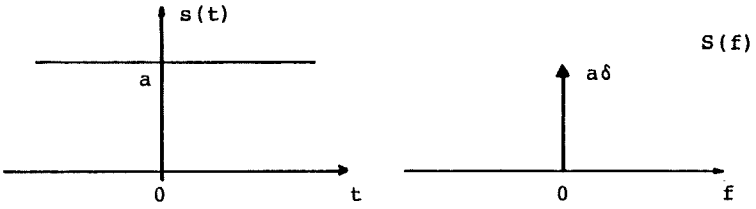
En δ-funktion, der har massen a koncentreret i punktet b, bliver

$$a\,\delta(t-b)\ .$$

Ved hjælp af forskellige former for grænseovergange (f. eks. på funktionerne i eksempel 9.7) kan man få en intuitiv forståelse for følgende relation

$$s(t) \equiv a \Rightarrow S(f) = a\,\delta(f)$$

d.v.s. den Fourier-transformerede til en konstant er lig konstanten gange en δ-funktion.



Vi fortsætter nu rækken af eksempler.

Eksempel 9.8 Hvis s er Heavyside-funktionen, i.e.

$$s(t) = \frac{1}{2}\, I_{[0,\infty[}(t)\ ,$$

bliver

$$S(f) = \frac{1}{2}\,\delta(f) + \frac{1}{i2\pi f}$$

□

Eksempel 9.9 Vi betragter nu den Fourier-transformerede til en δ-funktion, d.v.s.

$$s(t) = \delta(t)$$

Af definitionsligningen (4) fås umiddelbart

$$S(f) = 1 \ .$$

Hvis

$$s_\beta(t) = \delta(t-\beta) \ ,$$

bliver

$$S_\beta(f) = e^{-i2\pi f\beta} \ .$$

□

Eksempel 9.10 Vi betragter en cosinus-svingning med periode Δ, der er skåret af ved $\frac{T}{2}$, i.e.

$$s(t) = a \cos \frac{2\pi t}{\Delta} \, I_{[-\frac{T}{2}, \frac{T}{2}]}(t)$$

d.v.s., at

$$S(f) = \frac{a}{2} \left\{ T \frac{\sin \pi T(f-\frac{1}{\Delta})}{\pi T(f-\frac{1}{\Delta})} + T \frac{\sin \pi T(f+\frac{1}{\Delta})}{\pi T(f+\frac{1}{\Delta})} \right\}$$

□

Ved grænseovergange i dette eksempel bringes man til at indse

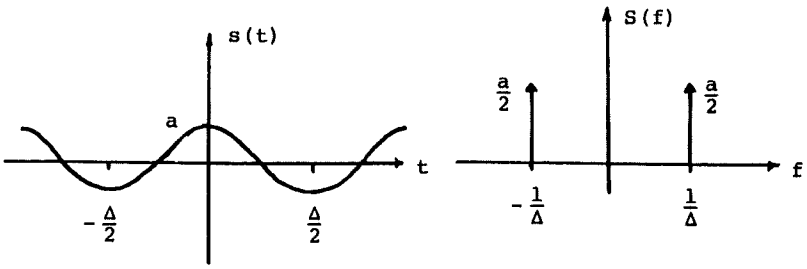
Eksempel 9.11 Cosinussvingningen med periode Δ (i.e. frekvens $\frac{1}{\Delta}$), i.e.

$$s(t) = a \cos \frac{2\pi t}{\Delta} \ ,$$

har den Fourier-transformerede

$$S(f) = \frac{a}{2}\left\{ \delta\left(f-\frac{1}{\Delta}\right) + \delta\left(f+\frac{1}{\Delta}\right) \right\} .$$

Graferne er



Eksempel 9.12 Betragter vi et signal med perioden Δ, d.v.s. et signal, der kan skrives

$$s(t) = \sum_{m=-\infty}^{\infty} S_m \, e^{i2\pi mt/\Delta} ,$$

fås

$$S(f) = \sum_{m=-\infty}^{\infty} S_m \, \delta\left(f-\frac{m}{\Delta}\right) ,$$

d.v.s. den Fourier-transformerede er et såkaldt "tog" af δ-funktioner.

□

Eksempel 9.13 Betragter vi omvendt et tog af δ-funktioner som signal, i.e.

$$s(t) = \sum_{n=-\infty}^{\infty} \delta(t-n\Delta) ,$$

fås

$$S(f) = \frac{1}{\Delta} \sum_{n=-\infty}^{\infty} \delta\left(f - \frac{n}{\Delta}\right) ,$$

altså igen et tog (se f. eks. Papoulis (1962) p. 44).

□

Efter denne lange række af eksempler nævner vi et par regneregler for Fourier integraler. Vi har

Sætning 9.2 (Nulpunkts- og skalaændring) Hvis signalet s(t) har den Fourier-transformerede S(f), vil

$s(at + b)$

have Fourier-transformationen

$\frac{1}{\lvert a \rvert} e^{i2\pi fb/a} S\left(\frac{f}{a}\right)$

Sætning 9.3 (Differentiation) Hvis s(t) har den Fourier-transformerede s(f), vil den m'te afledede

$s^{(m)}(t)$

have den Fourier transformerede

$(i2\pi f)^m S(f)$.

Sætning 9.4 (Symmetri) Hvis S(f) er den Fourier-transformerede af s(t), er s(f) den Fourier-transformerede af S(-t).

Beviser Ved at anvende definitionen direkte volder godtgørelsen af sætningerne 9.2 - 9.4 ingen vanskeligheder.

Sætning 9.5 Lad s(t) og w(t) have de Fourier-transformerede S(f) og W(f). Den Fourier-transformerede til produktet

$s(t)\ w(t)$

er da

$$S_p(f) = \int_{-\infty}^{\infty} S(g)\, W(f-g)\, dg = \int_{-\infty}^{\infty} W(g)\, S(f-g)\, dg \ ,$$

d.v.s. foldningsintegralet af S og W.

Omvendt gælder

Sætning 9.6 Lad s, w, S og W være som i sætning 9.5. Signalet svarende til den Fourier-transformerede

$$S(f)\, W(f)$$

er

$$s_q(t) = \int_{-\infty}^{\infty} w(u)s(t-u)\, du = \int_{-\infty}^{\infty} s(u)w(t-u)\, du \ ,$$

d.v.s. igen foldningsintegralet.

Bevis Beviset for sætningerne 9.5 og 9.6 er ligefremt. □

Vi kan kombinere de to sætninger i nedenstående corollar.

Corollar Ved overgangen mellem signal og Fourier-transformeret svarer multiplikation i tidsdomænet til foldning i frekvensdomænet og omvendt.

9.1.2 Forskydningsoperatorer

Vi indfører nu en række operatorer, der virker på følger som

$$\ldots, z_t, z_{t+1}, \ldots \ ,$$

noget vi senere vil betegne en tidsrække. Operatorerne vil svare til differentiation med hensyn til tiden i det tilfælde,

hvor der ikke observeres et signal til diskrete tidspunkter $\cdots,t,t+1,\cdots$, men derimod registreres kontinuerligt. Det er selvfølgelig helt betydningsløst, at vi har valgt tidsafstanden 1 mellem registreringerne.

Visse af definitionerne vil være af rentformel karakter (e.g. vil vi betragte uendelige summer af tal uden at bekymre os, om leddene går mod 0 eller ej. Den grundige læser kan enten søge en mere stringent fremstilling i litteraturen - f. eks. angående såkaldte polynomoperatorer -, eller han kan trøste sig med, at alle i praksis forekommende tidsrækker er 0 fra et vist trin at regne).

Den bagudrettede forskydningsoperator B defineres ved

$$B\, z_t = z_{t-1} .$$

I økonometrisk litteratur benævnes B ofte L (for lag-operator. Lag er engelsk for tidsafstand, forsinkelse).

B anvendt j gange bliver

$$B^j z_t = z_{t-j} .$$

Tilsvarende haves den fremadrettede forskydningsoperator F, der defineres ved

$$F\, z_t = z_{t+1}$$

og

$$F^j z_t = z_{t+j} .$$

Den bagudrettede differensoperator eller blot differensoperatoren ∇ defineres ved

$$\nabla\, z_t = z_t - z_{t-1} .$$

Skrives enhedsoperatoren (den identiske afbildning) 1 (d.v.s. $1\, z_t = z_t$), har vi

$$\nabla z_t = 1\, z_t - B\, z_t = (1-B)\, z_t ,$$

d.v.s.

$$\boxed{\nabla = 1 - B .}$$

Endelig haves summationsoperatoren S:

$$\boxed{S\, z_t = z_t + z_{t-1} + z_{t-2} + \cdots .}$$

De anførte operatorer er lineære, d.v.s.

$$H(az_t + by_t) = a(Hz_t) + b(Hy_t),$$

hvor H er en vilkårlig operator og a,b reelle eller komplekse tal. Endvidere er de kommutative, d.v.s. for vilkårlige operatorer H_1 og H_2 gælder

$$H_1 H_2 z_t = H_2 H_1 z_t .$$

Dette kan udnyttes, når man søger formen for f. eks. sammensatte operatorer. Vi anfører

Eksempel 9.14 Vi vil finde et direkte udtryk for en andenordens differens $\nabla^2 z_t$. Vi har

$$\nabla^2 = (1-B)^2 = 1 - 2B + B^2$$

d.v.s.

$$\nabla^2 z_t = z_t - 2z_{t-1} + z_{t-2} .$$

Denne formel verificeres også let direkte.

□

Har vi en potensrække - eventuelt med konvergensradius 0 -

$$a(z) = \sum_{\nu=0}^{\infty} a_\nu z^\nu$$

defineres for en operator H en ny operator a(H) ved

$$a(H)z_t = \sum_{\nu=0}^{\infty} a_\nu H^\nu z_t$$

eller kort

$$a(H) = \sum_{\nu=0}^{\infty} a_\nu H^\nu \ .$$

(Vi bemærker, at hvis alle a_ν er 0 fra et vist trin, er t et polynomium.)

Med denne definition har man - lidt løst formuleret - etableret en isomorfi mellem klassen af betragtede operatorer og mængden af potensrækker (evt. polynomier), således at der til produktet af to potensrækker svarer den sammensatte af to operatorer og til den inverse potensrække (hvis den eksisterer) den inverse operator. Vi formulerer resultaterne mere præcist i følgende sætninger og exempler.

Sætning 9.7 For en operator H betrages operatorerne

$$\phi(H) = \sum_{\nu=0}^{\infty} \phi_\nu H^\nu$$

$$\psi(H) = \sum_{\nu=0}^{\infty} \psi_\nu H^\nu$$

$$\pi(H) = \sum_{\nu=0}^{\infty} \pi_\nu H^\nu$$

og lad

$$\phi(H)\psi(H) = \pi(H) \ .$$

Da tilfredsstiller koefficienter ϕ_ν, ψ_ν og π_ν følgende ligningssystem

$$\begin{aligned}
\pi_0 &= \phi_0\psi_0 \\
\pi_1 &= \phi_0\psi_1 + \phi_1\psi_0 \\
&\vdots \\
\pi_j &= \phi_0\psi_j + \phi_1\psi_{j-1} + \cdots + \phi_{j-1}\psi_1 + \phi_j\psi_0 \\
&\vdots
\end{aligned}$$

Bevis Som for den tilsvarende sætning for potensrækker.

Corollar Vi har betegnelserne fra foregående sætning. Hvis $\psi = \phi^{-1}$ tilfredsstiller koefficienterne

$$\begin{aligned}
1 &= \phi_0\psi_0 \\
0 &= \phi_0\psi_1 + \phi_1\psi_0 \\
&\vdots \\
0 &= \phi_0\psi_j + \phi_1\psi_{j-1} + \cdots + \phi_{j-1}\psi_1 + \phi_j\psi_0
\end{aligned}$$

Bevis Fås trivielt af sætning 9.7 ved at sætte $\pi = 1$, dvs. den identiske afbildning.

Vi anfører nu et illustrativt eksempel.

Eksempel 9.15 Vi vil bestemme den inverse operator til differensoperatoren ∇. Vi har

$$\nabla = 1 - B$$

hvorfor

$$\begin{aligned}
\nabla^{-1} &= (1-B)^{-1} \\
&= 1 + B + B^2 + \cdots \\
&= S
\end{aligned}$$

d.v.s. ∇ og S er hinandens inverse.

□

9.1.3 Tidsinvariante, lineære systemer

Vi betragter nu systemer, som konverterer et input x(t) til et output y(t).

At et sådant system er lineært, vil sige, at, hvis

$$x_i(t) \to y_i(t) \quad , \; i=1,\cdots,n \; ,$$

da vil

$$\sum_{i=1}^{n} \alpha_i x_i(t) \to \sum_{i=1}^{n} \alpha_i y_i(t) \; .$$

At systemet er tidsinvariant betyder, at

$$x(t) \to y(t)$$

medfører

$$x(t-\tau) \to y(t-\tau) \quad , \quad \forall \tau \; .$$

Man kan da vise følgende

Sætning 9.8 For et lineært, tidsinvariant system findes der en funktion h, således at forbindelsen mellem input x og output y er givet ved foldningsintegralet mellem x og h, d.v.s.

$$y(t) = \int_{-\infty}^{\infty} h(u)x(t-u)\,du = \int_{-\infty}^{\infty} x(u)h(t-u)\,du$$

Bevis Forbigås.

Funktionen h kaldes impuls-responsfunktionen eller vægtfunktionen. Forklaringen på det første udtryk er, at h er outputtet, hvis inputtet er en δ-funktion, i.e.

$$y(t) = \int_{-\infty}^{\infty} \delta(u)h(t-u)\,du = h(t) \ .$$

Den Fourier-transformerede til output y(t) er ifølge sætning 9.6

$$Y(f) = H(f)X(f) \ ,$$

hvor H(f) og X(f) er de Fourier-transformerede til h(t) henholdsvis x(t).

H(f) kaldes frekvens-responsfunktionen. Splittes H(f) op i modulus og argument, d.v.s.

$$H(f) = |H(f)|e^{i\varphi(f)} = G(f)e^{i\varphi(f)} \ ,$$

kaldes G(f) forstærkningen (eng.: gain-function) og $\varphi(f)$ fase-funktionen.

Lineære systemer kan også beskrives ved hjælp af differentialligninger eller differensligninger. Vi formulerer nogle resultater i følgende to sætninger:

Sætning 9.9 Differentialligningen

$$\varphi_0\, y(t) - \varphi_1 \frac{dy(t)}{dt} - \cdots - \varphi_p \frac{d^p y(t)}{dt^p}$$

$$= \theta_0\, x(t-\tau) - \theta_1 \frac{dx(t-\tau)}{dt} - \cdots - \theta_q \frac{d^q x(t-\tau)}{dt^q}$$

angiver et lineært tidsinvariant system ved overgang fra input x(t) til output y(t). Frekvensresponsfunktionen for systemet er

$$H(f) = \frac{\theta_0 - \theta_1(i2\pi f) - \cdots - \theta_q(i2\pi f)^q}{\varphi_0 - \varphi_1(i2\pi f) - \cdots - \varphi_p(i2\pi f)^p} e^{-i2\pi f\tau}$$

Defineres polynomierne φ og θ ved

$$\varphi(z) = \varphi_0 - \varphi_1 z - \cdots - \varphi_p z^p$$

$$\theta(z) = \theta_0 - \theta_1 z - \cdots - \theta_q z^q$$

og sætter vi differentiationsoperatoren (med hensyn til t) lig D, kan differentialligningen skrives

$$\varphi(D)y(t) = \theta(D)x(t-\tau)$$

og frekvensresponsfunktionen bliver

$$H(f) = \frac{\theta(i2\pi f)}{\varphi(i2\pi f)} e^{-i2\pi f\tau}$$

Bevis Systemet bliver lineært, fordi differentialligningen er lineær, og tidsinvariant, fordi koefficienterne er konstante. Ved hjælp af sætning 9.8, p. 19, ses, at output $y(t)$ for input $x(t) = \exp(i2\pi ft)$ er $\exp(i2\pi ft)H(f)$. Indsættes denne relation i differentialligningen, fås udtrykket for $H(f)$ ved simple regninger.

Q.E.D.

Sætning 9.10 Differensligningen

$$y(t) - \varphi_1 y(t-\Delta) - \cdots - \varphi_p y(t-p\Delta) = \theta_0 x(t) - \theta_1 x(t-\Delta) - \cdots - \theta_q x(t-q\Delta)$$

angiver et lineært, tidsinvariant system ved overgang fra input $x(t)$ til output $y(t)$. Frekvensresponsfunktionen for systemet er

$$H(f) = \frac{\theta_0 - \theta_1 e^{-i2\pi f\Delta} - \cdots - \theta_q e^{-i2\pi fq\Delta}}{1 - \varphi_1 e^{-i2\pi f\Delta} - \cdots - \varphi_p e^{-i2\pi fp\Delta}}$$

Defineres polynomierne φ og θ ved

$$\varphi(z) = 1 - \varphi_1 z - \cdots - \varphi_p z^p$$

$$\theta(z) = \theta_0 - \theta_1 z - \cdots - \theta_q z^q$$

og indføres forskydningsoperatoren B ved $B\,z(t) = z(t-\Delta)$, jfr. p. 9.15, kan differensligningen skrives

$$\varphi(B)y(t) = \theta(B)x(t) ,$$

og frekvensresponsfunktionen bliver

$$H(f) = \frac{\theta(\exp(-i2\pi f\Delta))}{\varphi(\exp(-i2\pi f\Delta))}$$

Bevis Analogt til beviset for den forrige sætning med hensyn til linearitet og tidsinvarians. H(f) findes ved at tage Fourier-integralet på begge sider af lighedstegnet og så benytte sætning 9.2, p. 9.13.

Q.E.D.

En vigtig egenskab ved et lineært system er spørgsmålet om stabilitet.

Definition 9.1 Vi siger, at et system er stabilt, hvis et begrænset input giver et begrænset output.

I nedenstående sætning skal vi anføre en række betingelser for, at nogle lineære systemer er stabile.

Sætning 9.11 En tilstrækkelig betingelse for, at et lineært system med impulsresponsfunktion h(t) er stabilt, er

$$\int_{-\infty}^{\infty} |h(u)|du < \infty .$$

Systemet beskrevet ved differentialligningen i sætning 9.9 er stabilt, hvis rødderne i det karakteristiske polynomium

$$\varphi(z) = \varphi_0 - \varphi_1 z - \cdots - \varphi_p z^p$$

har negative realdele.

Systemet beskrevet ved differensligningen i sætning 9.10 er stabilt, hvis rødderne i det karakteristiske polynomium

$$\varphi(z) = \varphi_0 - \varphi_1 z - \cdots - \varphi_p z^p$$

ligger uden for enhedscirklen.

Bevis Forbigås.

En triviel, men nyttig bemærkning om ikke-vekselvirkende lineære systemer er, at hvis de anbringes i serie, fås igen et lineært system. Frekvensresponsfunktionen for det resulterende system er simpelt hen produktet af de enkelte systemers frekvensresponsfunktioner.

En anden bemærkning, vi må gøre, er, at for et fysisk system beskrevet ved

$$y(t) = \int_{-\infty}^{\infty} h(u)x(t-u)\,du ,$$

må vi kræve, at vægtfunktionen $h(u)$ er nul for negative værdier af u. Thi ellers ville output $y(t)$ afhænge af kommende værdier af input, hvilket selvsagt ikke er rimeligt. Vi kan da skrive

$$y(t) = \int_{0}^{\infty} h(u)x(t-u)\,du .$$

Betingelsen

$$h(u) = 0 \quad \text{for} \quad u < 0 \ ,$$

kaldes betingelsen om fysisk realiserbarhed.

9.1.4 Sampling-problemet

I dette lille afsnit skal vi betragte nogle af de vanskeligheder, man kan komme ud for, når man sampler i et kontinuert signal.

Det første problem, man møder, er selvsagt, at man kun kan betragte et signal i et endeligt interval $\left[-\frac{T}{2}, \frac{T}{2}\right]$.

Kaldes signalet $s(t)$, og indfører vi

$$w(t) = I_{[-\frac{T}{2}, \frac{T}{2}]}(t) \ ,$$

ser vi, at effekten af at betragte signalet i perioden $-\frac{T}{2} \leq t \leq \frac{T}{2}$ er at multiplicere signalet $s(t)$ med data-vinduet $w(t)$, d.v.s. det signal, vi måler, er

$$s_T(t) = s(t)w(t) \ .$$

Ifølge sætning 9.5 bliver den Fourier-transformerede til det aflæste signal

$$\boxed{S_T(f) = \int_{-\infty}^{\infty} S(g)W(f-g)\, dg \ ,}$$

hvor (ifølge eksempel 9.7).

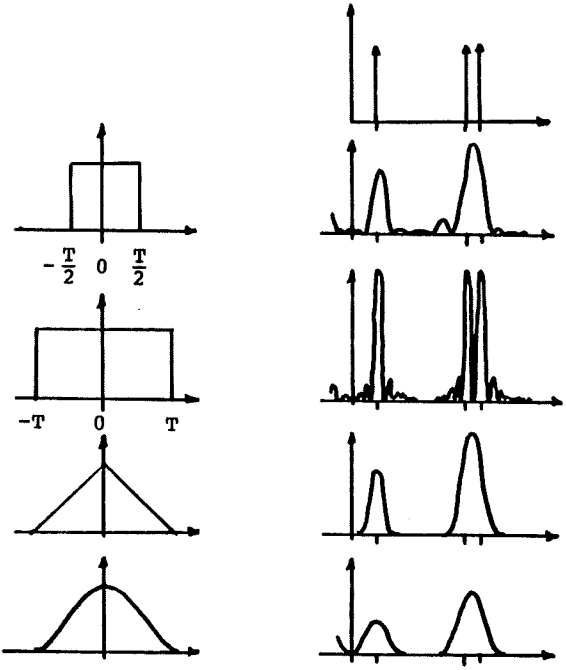
$$W(f) = T \frac{\sin \pi fT}{\pi fT} \ .$$

W kaldes spektral-vinduet.

Det er selvsagt ikke nødvendigt at anvende netop det anførte data-vindue. Ethvert rimeligt data-vindue vil give et spektral-vindue med en spids i 0, og som er forsvindende for store fre-

kvenser, d.v.s. som i nogen grad ligner det optimale, nemlig en δ-funktion placeret i 0. (Dette ville medføre $S_T \equiv S$).

Betydningen af data-vinduets form er vist i næste tegning, som viser det målte signals Fourier-transformation ved anvendelsen af forskellige data-vinduer. Tegningen er taget fra Jenkins & Watts (1968).



Fourier-transformation af signal (3 δ-fkt)

Fourier-transformation af målt signal ved de respektive data-vinduer. (Afstanden mellem de to spidser til højre er 1/T).

De mange små spidser på nogle af de aflæste Fourier-transformationer skyldes de skarpe kanter i data-vinduerne.

Det fremgår, at man må betragte signalet i et tidsrum af længden 2T, hvis man ved hjælp af et rektangulært data-vindue ønsker at skelne mellem de to spidser, der har en afstand af 1/T, eller

- ækvivalent hermed - hvis man ønsker at kunne skelne mellem spidser i frekvenserne f_1 og f_2 ($f_1 < f_2$), må man mindst sample i et tidsrum af længden $T > 2/(f_2 - f_1)$.

Det andet problem, man møder ved sampling af et kontinuert signal, er, at man ofte kun kan betragte det til diskrete tidspunkter

$$- n\Delta, -(n-1)\Delta, \cdots, 0, \cdots, (n-1)\Delta, n\Delta, \cdots .$$

Det samplede signal kan da opfattes som resultatet af at multiplicere det oprindelige signal med et tog af δ-funktioner, nemlig

$$j(t) = \sum_{n=-\infty}^{\infty} \delta(t-n\Delta) .$$

Dette giver det impuls-modulerede signal

$$s_j(t) = s(t)j(t) .$$

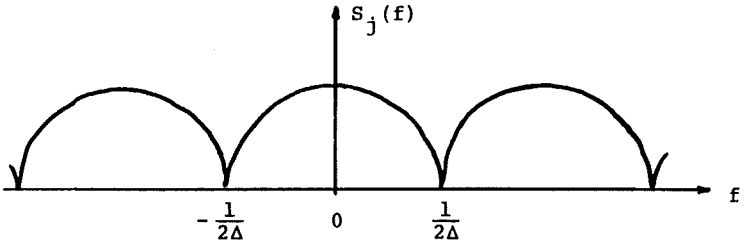
Ved hjælp af sætning 9.5 og eksempel 9.13 fås

$$\begin{aligned} S_j(f) &= \int_{-\infty}^{\infty} S(f-g)J(g)\, dg \\ &= \int_{-\infty}^{\infty} S(f-g)\, \frac{1}{\Delta} \sum_{n=-\infty}^{\infty} \delta(g-\tfrac{n}{\Delta})\, dg \\ &= \frac{1}{\Delta} \sum_{n=-\infty}^{\infty} S(f-\tfrac{n}{\Delta}) . \end{aligned}$$

Vi ser nu, at

$$S_j(f+\tfrac{k}{\Delta}) = \frac{1}{\Delta} \sum_{n=-\infty}^{\infty} S(f-\tfrac{n-k}{\Delta}) = S_j(f) ,$$

d.v.s. den Fourier-transformerede til det impuls-modulerede signal er periodisk med perioden $1/\Delta$.



Hvis S(f) derfor er 0 uden for intervallet

$$I_\Delta = \left[-\frac{1}{2\Delta}\ ,\ \frac{1}{2\Delta}\right] ,$$

er $S_j(f)$ blot en periodisk version af S(f). Vi kan derfor finde S(f) ud fra $S_j(f)$ ved en simpel multiplikation med

$$H(f) = \Delta\ I_{[-\frac{1}{2\Delta},\ \frac{1}{2\Delta}]}\ (f) ,$$

eller

$$S(f) = H(f)S_j(f) .$$

Ved hjælp af foldningssætningen fås

$$s(t) = \int_{-\infty}^{\infty} \frac{\sin(\pi u/\Delta)}{\pi u/\Delta}\ s_j(t-u)\ du .$$

Hvis S(f) derimod ikke er forsvindende uden for I_Δ, optræder der frekvenser, der er større end $\frac{1}{2\Delta}$ i $S_i(f)$, og vi kan ikke udskille disse. Vi siger, de er <u>aliased</u> eller <u>foldet</u> med de lavere frekvenser.

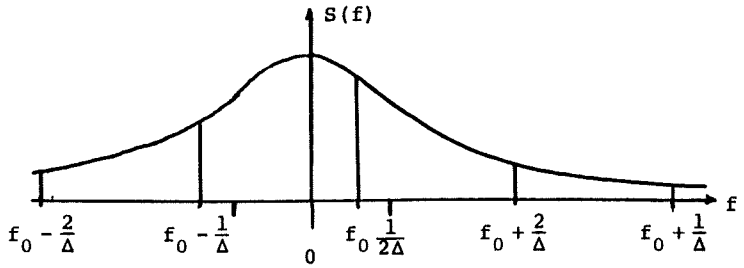
Vi kan da anvende omskrivningen

$$\Delta\ S_j(f) = S(f) + \sum_{n=1}^{\infty} [S(f+\tfrac{n}{\Delta}) + S(f-\tfrac{n}{\Delta})] ,$$

i.e. vægten til frekvensen f bliver blandet sammen med vægtene til frekvenserne

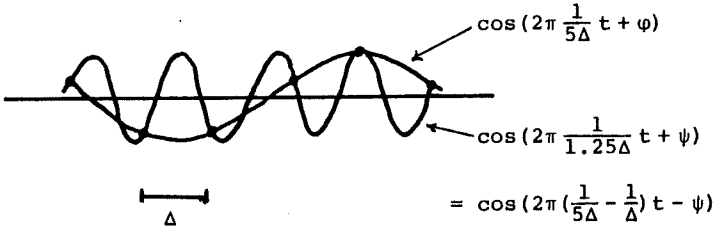
$$f \pm \frac{n}{\Delta} \quad , \quad n = 1,2,\cdots .$$

Dette er antydet i skitsen nedenfor.



Den største frekvens, der kan genfindes ved sampling med tidsintervallet Δ, er altså $\frac{1}{2\Delta}$, og den kaldes Nyqvist-frekvensen.

Det er ikke særlig vanskeligt at anskueliggøre aliasing-problemet grafisk i tidsdomænet:



Det er med den anførte værdi af Δ ikke muligt at skelne mellem de to cosinussvingninger, og derfor må et "estimat" over koefficienterne til den langsomtsvingende og til den hurtigtsvingende blive blandet sammen. Det skal lige anføres, at φ og ψ er faser, der her er uden betydning.

Vi kan her også antyde sammenhængen med definitionen af den diskrete Fourier-transformation. Ved rent formelle regninger fås for det kontinuerte signal s(t)

$$\Delta\, s_j(t) = \Delta\, s(t) j(t)$$

$$= \Delta \sum_{n=-\infty}^{\infty} s(t)\delta(t-n\Delta) \ .$$

Anvender vi nu definitionen på hvert led, fås

$$S(f) = \Delta\, S_j(f) = \Delta \sum_{n=-\infty}^{\infty} \int_{\infty}^{\infty} s(t)\delta(t-n\Delta)\ e^{-i2\pi ft}\ dt$$

$$= \Delta \sum_{n=-\infty}^{\infty} s(n\Delta)\ e^{-i2\pi fn\Delta} \ ,$$

hvor vi har forudsat, at S(f) er 0 uden for I_Δ. Den formel, vi her har opnået, viser altså, at den diskrete Fourier-transformation som defineret p. 9.3, kan opfattes som Fourier-integralet af et samplet signal fra en kontinuert funktion.

9.2 Stokastiske processer og deres momentfunktioner

Vi indleder med et afsnit med de basale definitioner.

9.2.1 Kort om tidsrækker og stokastiske processer

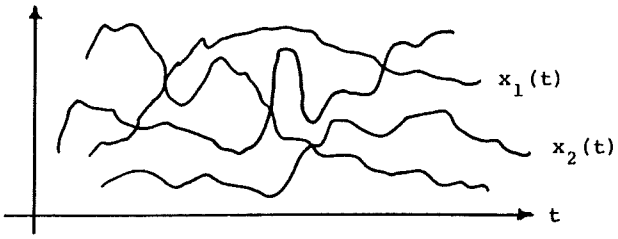
Ved en tidsrække vil vi her forstå en realisation x(t) af en stokastisk proces X(t). Den er altså en funktion af tiden, som afspejler stokastiske eller fluktuerende egenskaber, som kun kan beskrives ved sandsynlighedsteoretiske love.

Eksempler på sådanne er mangfoldige. Man kan f. eks. nævne de daglige børskurser for bestemte værdipapirer, vandstanden i en å, output fra et radarapparat etc., etc. Her er børskurserne et eksempel på en diskret tidsrække, idet der kun foreligger daglige værdier. Vandstandsmålingerne er til gengæld eksempel på en

kontinuert tidsrække, idet vi kan bestemme højde (f. eks. ved hjælp af en skriver) til et vilkårligt tidspunkt.

Som antydet ovenfor, må en analyse af en tidsrække derfor udføres som en undersøgelse af den underliggende stokastiske proces.

Mængden af mulige udfald af den stokastiske proces X(t) kaldes ensemblet hørende til processen.



De(n) tidsrækker, vi skal analysere må betragtes som et tilfældigt udtag af denne mængde. Oftest vil man kun have en enkelt realisation, og det vil da ikke være muligt at udtale sig særlig meget om processen, med mindre vi kan forudsætte, at en række egenskaber er kendte (situationen svarer fuldstændigt til, at man kun observerer et enkelt udfald af en stokastisk variabel X og på basis af denne ønsker at udtale sig om fordelingen af denne).

Vi skal nu give en kort oversigt over de vigtigste begreber, man tager i anvendelse ved beskrivelsen af en stokastisk proces.

Fordelingen af processen $X(t), t \in I$, (hvor $I \subseteq R$) beskrives ved alle endeligt-dimensionale frekvensfunktioner

$$f_{X(t_1),\ldots,X(t_n)}(x_1,\ldots,x_n) \ ,$$

hvor $t_1,\ldots,t_n \in I$ og $n \in N$.

Middelværdifunktionen for processen

$$\mu(t) = E(X(t)) = \begin{cases} \int_{-\infty}^{\infty} x f_{X(t)}(x)\,dx \\ \sum_{x} x f_{X(t)}(x) \end{cases} ,$$

hvor integraludtrykket anvendes, hvis X(t) er kontinuert fordelt, og summationsudtrykket, hvis den er diskret fordelt.

Tilsvarende defineres variansfunktionen som den afbildning, der til ethvert tidspunkt angiver variansen af X(t), i.e.

$$\sigma^2(t) = V(X(t)) = E([X(t)-\mu(t)]^2) .$$

Autokovariansfunktionen $\gamma(t_1,t_2)$ angiver kovariansen mellem $X(t_1)$ og $X(t_2)$, i.e.

$$\gamma_{XX}(t_1,t_2) = \gamma(t_1,t_2) = \mathrm{Cov}(X(t_1),X(t_2)) = E\Big([X(t_1)-\mu(t_1)][X(t_2)-\mu(t_2)]\Big).$$

Ofte betragtes autokorrelationsfunktionen

$$\rho_{XX}(t_1,t_2) = \rho(t_1,t_2) = \frac{\gamma(t_1,t_2)}{\sigma(t_1)\sigma(t_2)} ,$$

d.v.s. korrelationen mellem de variable målt til to (forskellige) tidspunkter.

Bemærkning I ovenstående har vi ikke taget hensyn til en række (tekniske) problemer vedrørende eksistens af de forskellige begreber som middelværdier og varianser etc. En sådan diskussion ligger uden for rammerne af denne fremstilling, og læseren må henvises til den righoldige litteratur om emnet.

Vi vil i det følgende især beskæftige os med processer, der i en passende forstand er tidsinvariante. Dette leder os til definitionen af stationære processer. Vi har først

Definition 9.2 Vi siger, at en proces $X(t), t \in I$, er strengt stationær, hvis alle de endeligt dimensionale fordelinger er invariante over for tidsforskydninger, d.v.s. hvis

$$\forall n \; \forall t_1,\ldots,t_n \; \forall h : f_{X(t_1),\ldots,X(t_n)}(x_1,\ldots,x_n) = f_{X(t_1+h),\ldots X(t_n+h)}(x_1,\ldots,x_n) .$$

Af dette udledes specielt, at

$$\forall t \; \forall h : \mu(t+h) = \mu(t) = \mu ,$$

d.v.s. middelværdifunktionen er konstant. Tilsvarende fås, at variansfunktionen er konstant, i.e.

$$\forall t \; \forall h : \sigma^2(t) = \sigma^2(t+h) = \sigma^2$$

For autokovariansen fås generelt

$$\begin{aligned} \gamma(t_1,t_2) &= \mathrm{Cov}\,(X(t_1),\, X(t_2)) \\ &= \mathrm{Cov}\,(X(t_1),\, X(t_1+(t_2-t_1))) \\ &= \mathrm{Cov}\,(X(t),\, X(t+u)) , \end{aligned}$$

hvor $t = t_1$ og $u = t_2 - t_1$. Hvis processen er strengt stationær, er denne imidlertid kun en funktion af u, og vi kan derfor skrive

$$\gamma(u) = \mathrm{Cov}\,(X(t),\, X(t+u))$$

for et vilkårligt t.

For autokorrelationen fås analogt

$$\rho(u) = \mathrm{Cor}\,(X(t),\, X(t+u)) = \frac{\gamma(u)}{\gamma(0)} .$$

Det skal bemærkes, at såvel γ som ρ er symmetriske funktioner, i.e.

$$\gamma(u) = \gamma(-u)$$

og

$$\rho(u) = \rho(-u) \ .$$

Det er især de ovennævnte egenskaber ved middelværdi-, varians-, autokovarians- og autokorrelationsfunktionen, som er af interesse ved analysen af tidsrækker. Vi indfører derfor et nyt begreb i

Definition 9.3 En proces $X(t)$, $t \in I$, er svagt stationær af orden k, hvis alle momenter af indtil k'te orden kun afhænger af tidsdifferenser. En svagt stationær proces af orden 2 kaldes blot svagt stationær.

Bemærkning Det vil formentlig være hensigtmæssigt at henlede læserens opmærksomhed på, at en vilkårlig funktion ikke altid kan optræde som autokovariansfunktion for en stationær stokastisk proces. Der gælder jo

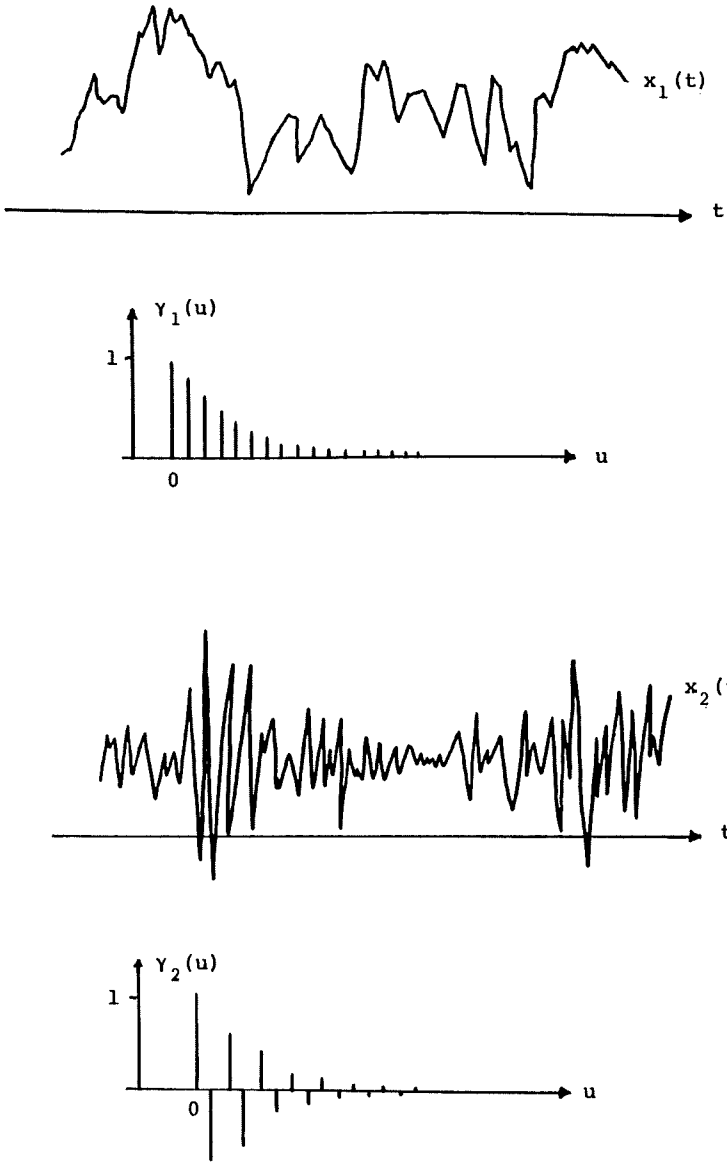
$$D\begin{pmatrix} X(t_1) \\ \vdots \\ X(t_n) \end{pmatrix} = \begin{pmatrix} \mathrm{Cov}(X(t_1),X(t_1)) & \dots & \mathrm{Cov}(X(t_1),X(t_n)) \\ \vdots & & \vdots \\ \mathrm{Cov}(X(t_n),X(t_1)) & \dots & \mathrm{Cov}(X(t_n),X(t_n)) \end{pmatrix}$$

$$= \begin{pmatrix} \gamma(0) & \dots & \gamma(t_n-t_1) \\ \vdots & & \vdots \\ \gamma(t_n-t_1) & \dots & \gamma(0) \end{pmatrix},$$

og denne matrix er positivt (semi-)definit. Dette lægger selvsagt bånd på γ.

□

I vedstående grafer er vist realisationer af to stokastiske processer (simulerede) og deres autokorrelationsfunktioner. Man bemærker de meget karakteristiske forskelle mellem tidsrækkerne.



Realiserede udfald af to stokastiske processer med tilhørende autokovariansfunktioner (efter Nelson (1973)).

Den første udviser kun meget langsomme svingninger. Hvis en observation er stor, er naboobservationerne det oftest også. Dette afspejles også i autokorrelationsfunktionen, der antager lutter positive værdier. Eksempelvis ses, at $\rho(1) \sim 0.8$, d.v.s. at korrelationen mellem to naboobservationer er ca. 0.8.

Omvendt med række nr. 2. Den fluktuerer meget voldsomt. Hvis en observation antager f. eks. en lille værdi, er naboobservationerne så at sige uden undtagelse store. Til gengæld er observationer 2 tidsenheder borte igen små, etc. Dette afspejles på nydelig vis i autokorrelationsfunktionen, der har skiftende fortegn.

Vi skal dernæst anføre nogle eksempler på stationære processer, som er ganske illustrative.

Eksempel 9.16 (Sinusproces). Hvis X og Y er uafhængige N(0,1)-fordelte stokastiske variable, vil den stokastiske proces defineret ved

$$Z_t = X \cos \tfrac{\pi}{3}t + Y \sin \tfrac{\pi}{3}t$$

være en stærkt stationær proces. Bruger vi de sædvanlige additionsformler for trigonometriske funktioner, kan vi skrive

$$Z_t = \sqrt{X^2+Y^2}\left[\frac{X}{\sqrt{X^2+Y^2}} \cos \tfrac{\pi}{3}t + \frac{Y}{\sqrt{X^2+Y^2}} \sin \tfrac{\pi}{3}t\right]$$

$$= R \cos \left(\tfrac{\pi}{3}t-\Phi\right) ,$$

hvor

$$R^2 = X^2 + Y^2 \in \chi^2(2) ,$$

og

$$\Phi = \operatorname{Arccos} \frac{X}{\sqrt{X^2+Y^2}} \in U(0,2\pi) .$$

Processen er altså en cosinussvingning med en $\chi(2)$-fordelt amplitude R, en periode på 6 med en tilfældig (rektangulært for-

delt over $[0,2\pi]$) faseforskydning Φ. En enkelt realisation består altså af en ren cosinussvingning med en given periode på 6. Amplituden og faseforskydningen varierer fra realisation til realisation.

Momentfunktionerne bliver

$$E(Z_t) = 0 \cdot \cos \tfrac{\pi}{3}t + 0 \cdot \sin \tfrac{\pi}{3}t = 0$$

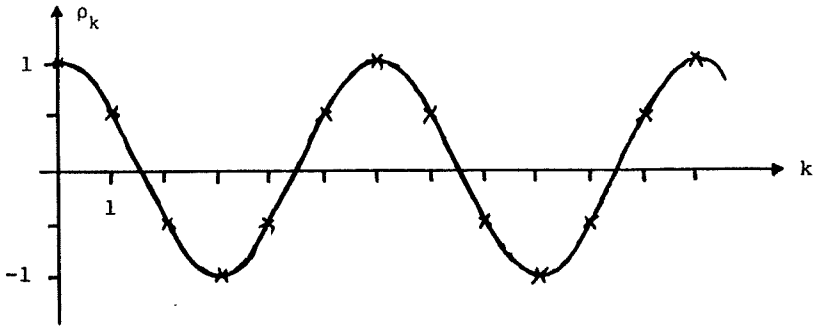
$$V(Z_t) = 1 \cdot \cos^2 \tfrac{\pi}{3}t + 1 \cdot \sin \tfrac{\pi}{3}t = 1$$

$$\begin{aligned}
\mathrm{Cov}(Z_t, Z_{t+k}) &= E\{X \cos \tfrac{\pi}{3}t + Y \sin \tfrac{\pi}{3}t\} \times \{X \cos \tfrac{\pi}{3}(t+k) + Y \sin \tfrac{\pi}{3}(t+k)\} \\
&= E\{X^2 \cos \tfrac{\pi}{3}t \cos \tfrac{\pi}{3}(t+k)\} + E\{Y^2 \sin \tfrac{\pi}{3}t \sin \tfrac{\pi}{3}(t+k)\} \\
&\qquad + E\{XY \cdot \text{konstant}\} \\
&= \cos \tfrac{\pi}{3}t \cos \tfrac{\pi}{3}(t+k) + \sin \tfrac{\pi}{3}t \sin \tfrac{\pi}{3}(t+k) \\
&= \cos \left[\tfrac{\pi}{3}t - \tfrac{\pi}{3}(t+k)\right] \\
&= \cos \tfrac{\pi}{3}k \ .
\end{aligned}$$

Da disse alle er uafhængige af t, følger den svage stationaritet umiddelbart. Da en flerdimensional normalfordeling er karakteriseret ved middelværdi og dispersionsstruktur, er alle endeligtdimensionale fordelinger af Z_t normale og tidsforskydningsinvariante, og deraf følger den stærke stationaritet.

Autokorrelationsfunktionens graf er skitseret p. 9.37.

Bemærk, at værdien 1 antages for værdier større end et vilkårligt positivt tal. Dette indebærer bl.a., at man vil være i stand til at lave perfekte forudsigelser for en vilkårlig lang tidshorisont, når vi har observeret processen i et passende langt, endeligt tidsinterval (ikke chokerende, når vi erindrer, at et udfald af processen er en ren cosinussvingning).



□

Eksempel 9.17 (Skjult sinusproces). Lad A_t, $t = \cdots,-1,0,1,\cdots$, være uafhængige $N(0,\sigma^2)$-fordelte stokastiske variable og sæt

$$Y_t = Z_t + A_t ,$$

hvor Z_t er som i foregående eksempel. Vi forudsætter, at A 'erne også er uafhængige af X og Y og dermed af Z_t'erne. Vi får da

$$E(Y_t) = 0$$

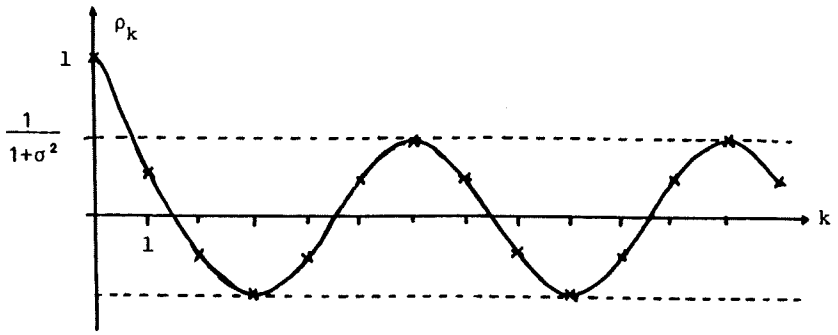
$$V(Y_t) = 1 + \sigma^2$$

$$\begin{aligned} \mathrm{Cov}(Y_t, Y_{t+k}) &= E\{(Z_t+A_t)(Z_{t+k}+A_{t+k})\} \\ &= E\{Z_t \cdot Z_{t+k}\} \\ &= \cos \tfrac{\pi}{3}k \ , \ k \neq 0 \ . \end{aligned}$$

Heraf følger stationariteten som i foregående eksempel. Vi finder autokorrelationsfunktionen til

$$\rho_k = \begin{cases} 1 & , \quad k = 0 \\ \dfrac{\cos \frac{\pi}{3}k}{1+\sigma^2} & , \quad k \neq 0 \end{cases} .$$

Grafen er anført på side 9.38.



Vi ser, at autokorrelationsfunktionen antager værdien $1/1+\sigma^2$ for værdier større end et vilkårligt tal. Dette modsvarer, at processen kan forudsiges med en forudsigelsesfejl, hvis varians ikke afhænger af tidshorisonten. (Forudsigelsen til tid t+k er $\hat{Y}_{t+k} = Z_{t+k}$. Fejlen er da A_t med varians σ^2). □

I næste eksempel er anført en simuleret realisation af Y_t (efter Wold (1965)).

Eksempel 9.18 (Forstyrret sinusproces). Vi betragter den homogene andenordens differensligning

$$x_t - \varphi x_{t-1} + \varphi^2 x_{t-2} = 0 \ . \qquad (7)$$

Løsningerne til denne er af formen (jfr. opgave 9.6)

$$x_t = a \cdot \varphi^t \cos(\theta t+\varepsilon) \ ,$$

hvor

$$\operatorname{tg} \theta = -\frac{\sqrt{4\varphi^2-\varphi^2}}{\varphi} = \begin{cases} -\sqrt{3} & , \ \varphi > 0 \\ \sqrt{3} & , \ \varphi < 0 \end{cases} ,$$

d.v.s

$$\theta = \begin{cases} -\frac{\pi}{3} & , \ \varphi > 0 \\ \frac{\pi}{3} & , \ \varphi < 0 \end{cases} .$$

Løsningerne til ligningen er derfor mere præcist

$$x_t = a\,\varphi^t \cos\left(\frac{\pi}{3}t+\varepsilon\right) ,$$

hvor a og ε er vilkårlige konstanter.

Betragter vi altså et system, hvor "tilstanden" x_t styres af differensligningen (7), vil x_t altså være en cosinussvingning med en periode på 6. Påføres systemet nu en forstyrrelse, således at tilstanden ikke styres af (7), men af (7) med højresiden erstattet af uafhængige normaltfordelte stokastiske variable A_t med varians σ_a^2 og forventningsværdi 0, bliver "tilstanden" også stokastisk og tilfredsstiller altså

$$X_t - \varphi X_{t-1} + \varphi^2 X_{t-2} = A_t \qquad (8)$$

Den fremkomne proces X_t benævnes en andenordens autoregressiv proces med parametre $(\varphi,-\varphi^2)$. Vi vil senere beskæftige os mere dybtgående med processer af denne type. Det vil der fremgå, at X_t er stationær (se p. 9.51). Ved at tage middelværdier på begge sider af lighedstegnet i (8) fås med $E(X_t) = \mu_t = \mu$

$$\mu - \varphi\,\mu + \varphi^2\,\mu = 0 ,$$

hvilket medfører

$$E(X_t) = 0 .$$

Vi vil nu bestemme autokorrelationsfunktionen ρ_k. Vi godtgør først, at den tilfredsstiller (7). Ved multiplikation på begge sider af lighedstegnet i definitionsligningen (8) med X_{t-k} fås

$$X_t\,X_{t-k} = \varphi\,X_{t-1}\,X_{t-k} - \varphi^2\,X_{t-2}\,X_{t-k} + A_t\,X_{t-k} .$$

Ved at tage forventningsværdier fås nu

$$\gamma_k = \varphi\,\gamma_{k-1} - \varphi^2\,\gamma_{k-2} \quad , \quad k \neq 0$$

$$\gamma_0 = \varphi\,\gamma_1 - \varphi^2\,\gamma_2 + \sigma_a^2 \quad ,$$

hvor $\sigma_a^2 = V(A_t) = E(A_t^2) = E(A_t\,Y_t)$. Heraf fås ved division med γ_0

$$\rho_k = \varphi\,\rho_{k-1} - \varphi^2\,\rho_{k-2}$$

$$\rho_0 = 1 \ .$$

Altså tilfredsstiller ρ_k (7), og vi må derfor have

$$\rho_k = a\,\varphi^k \cos\left(\tfrac{\pi}{3}k+\varepsilon\right) \ ,$$

hvor de arbitrære konstanter a og ε skal fastlægges. Ved at sætte k=1 i differensligningen fås

$$\rho_1 = \varphi\,\rho_0 - \varphi^2\,\rho_1$$

eller

$$\rho_1 = \frac{\varphi}{1+\varphi^2} \ .$$

Ved at indsætte 0 fås

$$\rho_0 = 1 = a \cos\varepsilon$$

eller

$$a = \frac{1}{\cos\varepsilon} \ .$$

Indsættes det fundne udtryk for ρ_1 og værdien for a i differensligningen fås

$$\frac{1}{\cos\ \varepsilon}\ \varphi\ \cos\left(\frac{\pi}{3}+\varepsilon\right)\ =\ \frac{\varphi}{1+\varphi^2}\ ,$$

eller

$$\cos\ \frac{\pi}{3}\ -\ \sin\ \frac{\pi}{3}\ \mathrm{tg}\ \varepsilon\ =\ \frac{1}{1+\varphi^2}\ ,$$

d.v.s.

$$\mathrm{tg}\ \varepsilon\ =\ -\frac{1}{\sqrt{3}}\ \frac{1-\varphi^2}{1+\varphi^2}$$

Med dette kan udtrykket for ρ_k omskrives til

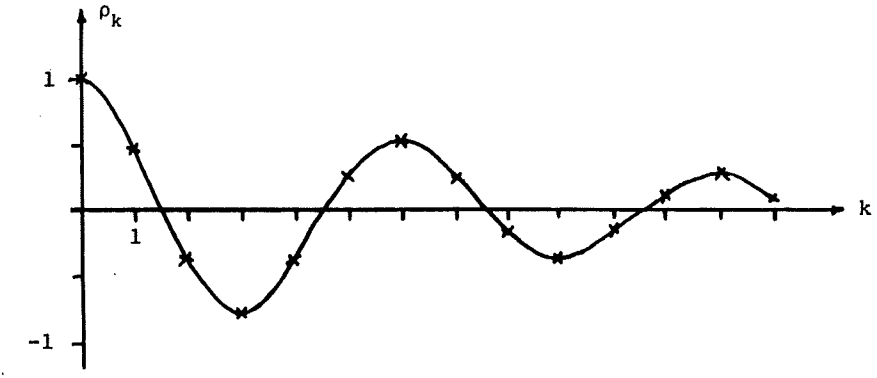
$$\rho_k\ =\ \varphi^k\ \frac{1}{\cos\ \varepsilon}\left[\cos\ \frac{\pi}{3}k\ \cos\ \varepsilon - \sin\ \frac{\pi}{3}k\ \sin\ \varepsilon\right]$$

$$=\ \varphi^k\left[\cos\ \frac{\pi}{3}k - \mathrm{tg}\ \varepsilon\ \sin\ \frac{\pi}{3}k\right]\ ,$$

d.v.s.

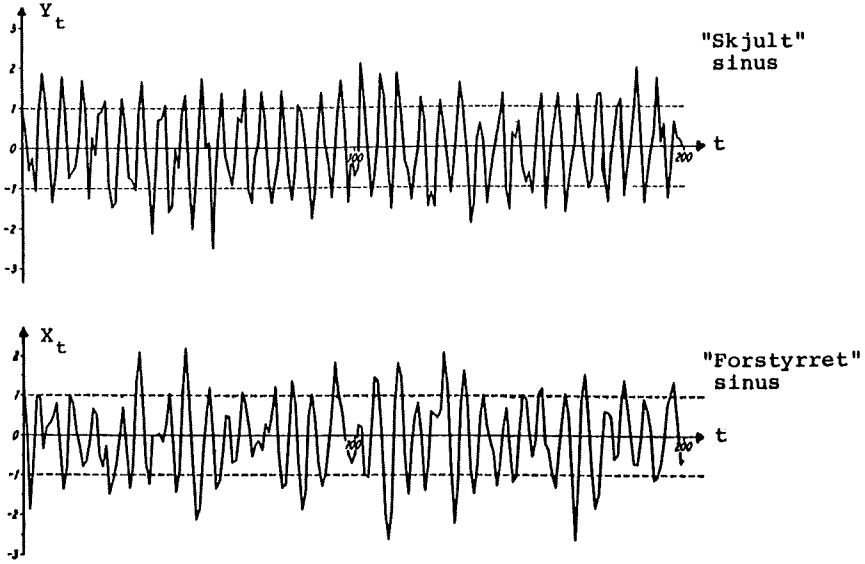
$$\rho_k\ =\ \varphi^k\left[\cos\ \frac{\pi}{3}k + \frac{1}{\sqrt{3}}\ \frac{1-\varphi^2}{1+\varphi^2}\ \sin\ \frac{\pi}{3}k\right]\ .$$

Grafen for ρ_k er i tilfældet $\varphi = 0.9$ skitseret nedenfor



Vi ser, at korrelationsfunktionen dør eksponentielt ud, og vi skal senere se, at det modsvares af, at det kun er muligt at lave den trivielle forudsigelse $\hat{X}_{t+k} \approx 0$ for tilstrækkeligt store værdier af k.

Vi bemærker, at den grundlæggende periode 6 går igen i de sidste eksempler. For at illustrere forskellen på processerne i de to sidste eksempler skal nedenfor anføres simulerede versioner af Y_t og X_t (for $\gamma = 0.9$) (efter Wold (1965)).



Vi bemærker, at perioden er helt fast for den skjulte sinus, hvorimod der fra tid til anden optræder tilfældige afvigelser i den forstyrrede sinus, således at afstanden mellem "toppe" afviger fra 6.

□

De foregående eksempler illustrerer nogle vigtige begreber inden for teorien om stationære processer, nemlig de såkaldte deterministiske processer (svarende til sinusprocessen), de såkaldte rent indeterministiske processer (svarende til den autoregressive proces eller den forstyrrede sinus) og de blandede processer (som f.eks. den skjulte sinusproces).

Uden at gå i detaljer med hensyn til de mere målteoretiske aspekter skal vi anføre følgende definitioner.

Definition 9.4 En stationær proces Z_t siges at være deterministisk, hvis den kan bringes på formen

$$Z_t = \sum_{i=1}^{n} (X_i \cos \lambda_i t + Y_i \sin \lambda_i t)$$

$$= \sum_{i=1}^{n} R_i \cos(\lambda_i t - \Phi_i) \quad ,$$

hvor X_i og Y_i er ukorrelerede stokastiske variable med middelværdier $E(X_i) = E(Y_i) = 0$ og varianser $V(X_i) = V(Y_i) = \sigma_i^2$. R_i og Φ_i er givet ved

$$R_i^2 = X_i^2 + Y_i^2$$

og

$$\text{tg } \Phi_i = \frac{Y_i}{X_i} \quad .$$

Definition 9.5 En stationær proces X_t siges at være (rent) indeterministisk, hvis den kan bringes på formen

$$X_t = A_t + \beta_1 A_{t-1} + \beta_2 A_{t-2} + \cdots \quad ,$$

hvor A_t'erne er ukorrelerede med middelværdi 0 og varianser σ^2, og hvor

$$\sum_{\nu=0}^{\infty} \beta_\nu^2 < \infty \quad .$$

Bemærkning I fortsættelse af eksemplerne skal her blot slås fast, at processen Z_t kan predikteres perfekt til tidspunkt t+k ved linearkombinationer af tidligere værdier $Z_t, Z_{t-1}, \cdots$, hvorimod processen X_t kun lader sig forudsige med en varians

$$E(X_{t+k} - \hat{X}_{t+k})^2 = \sigma^2(1 + \beta_1^2 + \cdots + \beta_{k-1}^2) \quad .$$

Her er

$$\hat{X}_{t+k} = \beta_k A_t + \beta_{k+1} A_{t-1} + \beta_{k+2} A_{t-2} + \cdots \quad ,$$

se f.eks. beviset for sætning 9.45, p. 9.158.

En hovedsætning om stationære processer er nu

Sætning 9.12 (Wolds prediktive dekomposition). En stationær proces Y_t kan spaltes i en sum

$$Y_t = Z_t + X_t \quad ,$$

hvor Z_t er deterministisk, og X_t er rent indeterministisk.

Bevis Forbigås, se f.eks. Anderson (1971).

Et andet fundamentalt resultat om stationære processer er den såkaldte spektralrepræsentation af en stationær proces. Resultatet her er, at en stationær proces kan approximeres med et vægtet integral af trigonometriske funktioner, hvor vægtene er stokastiske variable. Dette skal vi dog ikke komme yderligere ind på her, men f.eks. henvise til Anderson (1971).

9.2.2 Den diskrete lineære proces

Vi skal i dette afsnit beskæftige os med en meget vigtig type af processer, nemlig de såkaldte lineære processer. Disse processer kan opfattes som output fra et lineært system, hvor input er en såkaldt hvid støj A_t (eller en fuldstændig tilfældig proces A_t), d.v.s. en proces af ukorrelerede, identisk fordelte variable med middelværdi 0. Vi har altså

$$\mu_t = E(A_t) = 0$$

$$\sigma_t^2 = V(A_t) = E(A_t^2) = \sigma^2$$

$$\gamma_k = \mathrm{Cov}(A_t, A_{t+k}) = E(A_t \cdot A_{t+k}) = 0 \quad , \quad k \neq 0$$

for en (diskret) hvid støj A_t, $t = \cdots,-1,0,1,\cdots$.

Definition 9.6 Ved en lineær proces Z_t forstås en proces af formen

$$Z_t - \mu = \sum_{j=0}^{\infty} \psi_t \, A_{t-j} = A_t + \psi_1 \, A_{t-1} + \cdots \quad . \tag{9}$$

Defineres generelt $\tilde{Z}_t$ som Z_t-processen justeret for niveau, kan vi skrive

$$\tilde{Z}_t = \sum_{j=0}^{\infty} \psi_j \, B^j \, A_t = \psi(B) A_t \quad ,$$

hvor B er den sædvanlige forskydningsoperator, og

$$\psi(B) = 1 + \sum_{j=1}^{\infty} \psi_j \, B^j \quad .$$

ψ_j'erne svarer til impulsresponsfunktionen for det lineære system, der genererer output $\tilde{Z}_t$ ud fra input A_t .

Bemærkning A_t'erne benævnes ofte random shocks i den angelsaksiske litteratur.

Ved rekursivt at løse ligningerne (9) med hensyn til A_t'erne kan vi skrive $\tilde{Z}_t$ som en linearkombination af tidligere værdier af $\tilde{Z}_t$. Vi har

$$\tilde{Z}_{t-1} = A_{t-1} + \psi_1 \, A_{t-2} + \psi_2 \, A_{t-3} + \cdots$$

og dermed

$$A_{t-1} = \tilde{Z}_{t-1} - \psi_1 \, A_{t-2} - \psi_2 \, A_{t-3} - \cdots \quad ,$$

o.s.v. Dette giver

$$\begin{aligned} \tilde{Z}_t &= \psi_1 \, \tilde{Z}_{t-1} + (\psi_2 - \psi_1^2) A_{t-2} + (\psi_3 - \psi_1 \, \psi_2) A_{t-3} + \cdots + A_t \\ &= \psi_1 \, \tilde{Z}_{t-1} + (\psi_2 - \psi_1^2) \tilde{Z}_{t-2} + (\psi_3 - 2\psi_1 \, \psi_2 - \psi_3^3) A_{t-3} + \cdots + A_t \\ &\vdots \\ \tilde{Z}_t &\quad \pi_1 \, \tilde{Z}_{t-1} + \pi_2 \, \tilde{Z}_{t-2} + \cdots + A_t \quad . \end{aligned} \tag{10}$$

Vi anfører relationen mellem de to fremstillinger i

Sætning 9.13 Vi betragter processen $\tilde{Z}_t$ givet ved (9) eller den ækvivalente form (10). Indføres operatorerne

$$\psi(B) = 1 + \sum_{j=1}^{\infty} \psi_j B^j$$

$$\pi(B) = 1 - \sum_{j=1}^{\infty} \pi_j B^j$$

kan (9) og (10) skrives

$$\tilde{Z}_t = \psi(B) A_t \qquad (11)$$

$$\pi(B)\tilde{Z}_t = A_t \quad , \qquad (12)$$

og koefficienterne i de to fremstillinger tilfredsstiller ligningssystemet

$$\begin{aligned}
0 &= \psi_1 - \psi_0\pi_1 \\
0 &= \psi_2 - \psi_0\pi_2 - \psi_1\pi_1 \\
0 &= \psi_3 - \psi_0\pi_3 - \psi_1\pi_2 - \psi_2\pi_1 \\
&\vdots \\
0 &= \psi_j - \psi_0\pi_j - \psi_1\pi_{j-1} - \cdots - \psi_{j-1}\pi_1 \\
&\vdots
\end{aligned}$$

hvor

$$\psi_0 = 1 \quad .$$

Bevis Resultaterne (11) og (12) er trivielle. Ved at anvende $\psi(B)$ på begge sider af (12) og benytte (11) fås

$$\psi(B)\pi(B)\tilde{Z}(t) = \psi(B)A_t = \tilde{Z}_t \quad ,$$

d.v.s. $\psi(B)$ og $\pi(B)$ er hinandens inverse. Resultatet følger nu let af sætning 9.7.

Q.E.D.

Det kan vises, at processen er stationær, hvis ψ'erne går tilpas hurtigt mod 0, d.v.s. hvis meget fortidige værdier af den hvide støj ikke har for stor indflydelse. Mere præcist gælder

Sætning 9.14 Den lineære proces $\tilde{Z}_t = \psi(B)A_t$ er stationær, hvis potensrækken

$$\psi(z) = 1 + \sum_{j=1}^{\infty} \psi_j z^j$$

konvergerer for $|z| \leq 1$, $z \in C$.

Bevis Det detaljerede bevis forbigås, se f.eks. Grenander & Rosenblatt (1957). I det store og hele følger det dog af beviset for nedenstående sætning 9.15, hvor det godtgøres, at $\mathrm{Cov}(\tilde{Z}_t, \tilde{Z}_{t+k})$ er uafhængig af t.

Et helt analogt resultat om π-vægtene har vi i nedenstående definition (NB!)

Definition 9.7 Den lineære proces $\pi(B)\tilde{Z}_t = A_t$ er invertibel, hvis

$$\pi(z) = 1 - \sum_{j=1}^{\infty} \pi_j z^j$$

konvergerer for $|z| \leq 1$.

Bemærkning Betingelsen, at π_j'erne skal gå tilpas hurtigt mod 0, er en betingelse om, at fortidige observationer skal have en stærkt dalende indflydelse. Begrebet virker i sig selv meget fornuftigt; men der er især tale om en ad hoc definition, som skal sikre identificerbarheden af de såkaldte ARIMA-processer.

Vi anfører autokovariansen i

Sætning 9.15 Den lineære proces

$$\tilde{Z}_t = \psi(B)A_t = \sum_{j=0}^{\infty} \psi_j A_{t-j}$$

har autokovariansfunktion

$$\gamma_k = \sigma_a^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+k} \quad ,$$

hvor $\sigma_a^2 = V(A_t)$.

Bevis Helt ligefremt. Af definitionen følger

$$\gamma_k = E\left[\tilde{Z}_t \tilde{Z}_{t+k}\right]$$

$$= E\left[\sum_{j=0}^{\infty} \sum_{h=0}^{\infty} \psi_j \psi_h A_{t-j} A_{t+k-h}\right]$$

$$= \sigma_a^2 \sum_{j=0}^{\infty} \psi_j \psi_{j+k} \quad ,$$

da $E\left[A_{t-j} A_{t+k-h}\right] = 0$ for $t - j \neq t + k - h$.

Q.E.D.

Vi specialiserer nu til tilfælde, hvor ψ_j og π_j er nul fra et vist trin. Vi taler da om glidende gennemsnitsprocesser og autoregressive processer. Vi har

Definition 9.8 Den lineære proces bestemt ved

$$\tilde{Z}_t = A_t - \theta_1 A_{t-1} - \theta_2 A_{t-2} - \cdots - \theta_q A_{t-q} \quad ,$$

benævnes en glidende gennemsnitsproces af orden q eller kort en MA(q)-proces (MA for moving average). Defineres q'te grads polynomiet θ ved

$$\theta(z) = 1 - \theta_1 z - \theta_2 z^2 - \cdots - \theta_q z^q$$

kan vi også skrive

$$\tilde{Z}_t = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q) A_t = \theta(B) A_t \ .$$

For en sådan proces har vi

Sætning 9.16 En MA(q)-proces er altid stationær, men kun invertibel, hvis rødderne i ligningen $\theta(z) = 0$, d.v.s.

$$1 - \theta_1 z - \theta_2 z^2 - \cdots - \theta_q z^q = 0$$

ligger uden for enhedscirklen, eller ækvivalent hermed, hvis rødderne i

$$z^q - \theta_1 z^{q-1} - \cdots - \theta_q = 0$$

ligger inden i enhedscirklen. Begge ligninger kaldes processens karakteristiske ligning.

Bevis Stationariteten følger umiddelbart. Betingelsen om invertibilitet følger efter nogen regning af definitionen ved at betragte koefficienten i $\theta^{-1}(B)$.

Vi anfører momentfunktionerne for en MA(q)-proces i

Sætning 9.17 For en MA(q)-proces $\tilde{Z}_t = \theta(B) A_t$ er variansen

$$\sigma_z^2 = \gamma_0 = (1 + \theta_1^2 + \cdots + \theta_q^2)\sigma_a^2 \ ,$$

og autokorrelationen

$$\rho_k = \begin{cases} \dfrac{-\theta_k + \theta_1\ \theta_{k+1} + \cdots + \theta_{q-k}\ \theta_q}{1 + \theta_1^2 + \cdots + \theta_q^2} & , \ k = 1, \cdots, q \\ 0 & , \ k > q \end{cases}$$

Bevis Følger direkte af sætning 9.15.

Eksempel 9.19 I nedenstående graf er vist realisationer af en hvid støj A_t med $V(A_t) = 1$ og af 2 MA(1)-processer, nemlig

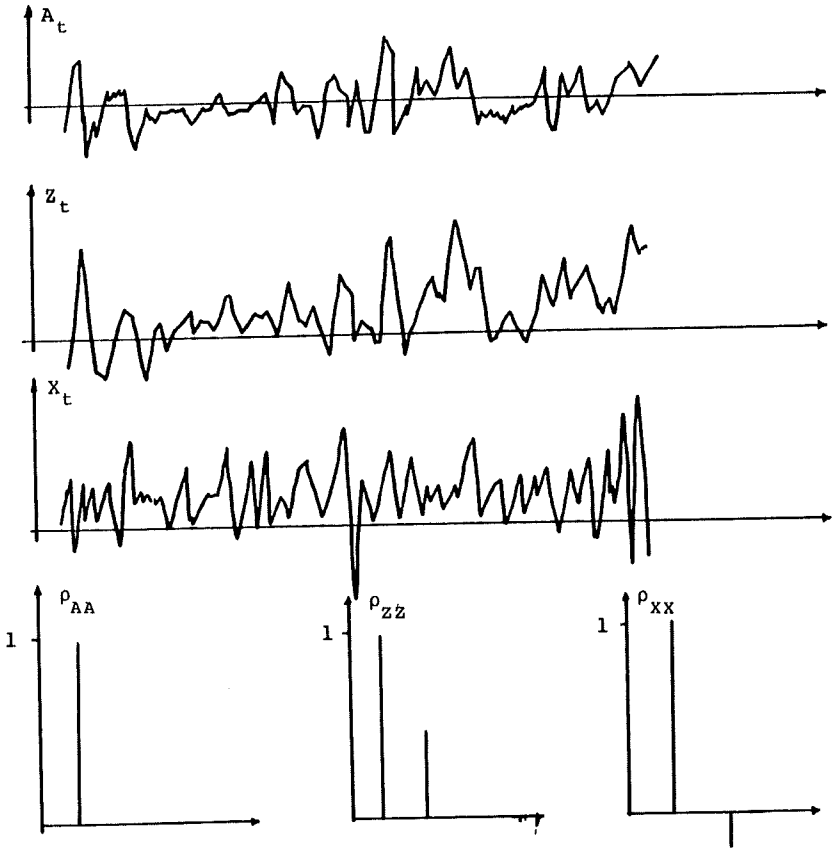
$$Z_t = 1 + A_t + 0.9\ A_{t-1}$$

og

$$X_t = 1 + A_t - 0.9\ A_{t-1} \ .$$

Endvidere er korrelationsfunktionerne angivet.

Man bemærker, at Z_t-processen svinger langsommere og X_t-processen hurtigere end den fuldstændig tilfældige proces A_t. Forklaringen findes let i autokorrelationsfunktionerne, idet $\rho_{ZZ}(1) > 0$, og $\rho_{XX}(1) < 0$.



Hvid støj og to MA(1)-processer samt disses autokorrelationsfunktioner. Efter Nelson (1973).

□

Vi betragter dernæst den p'te ordens autoregressive proces jfr. eksempel 9.18.

Definition 9.9 En p'te ordens autoregressiv proces, eller kort en AR(p)-proces, er en lineær proces af formen

$$\tilde{Z}_t = \phi_1 \tilde{Z}_{t-1} + \phi_2 \tilde{Z}_{t-2} + \cdots + \phi_p \tilde{Z}_{t-p} + A_t \qquad (13)$$

eller

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)\, \tilde{Z}_t = A_t$$

eller

$$\phi(B)\, \tilde{Z}_t = A_t \,, \qquad (14)$$

hvor ϕ altså er et polynomium af p'te grad.

En autoregressiv proces er trivielt invertibel. Stationaritetskravet bliver

Sætning 9.18 Den autoregressive proces (13) er stationær, såfremt alle rødder i $\phi(z) = 0$, d.v.s.

$$1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p = 0 \,,$$

ligger uden for enhedscirklen eller ækvivalent hermed, hvis alle rødder i

$$z^p - \phi_1 z^{p-1} - \cdots - \phi_p = 0$$

ligger inden i enhedscirklen.

Bevis Forbigås.

Bemærkning Begge de anførte ligninger omtales i litteraturen som den karakteristiske ligning for processen. Man må derfor ved referencer o.l. være opmærksom på, hvilken af de to, der er tale om.

I eksempel 9.18, p. 9.38, fandt vi autokorrelationsfunktionen for en AR(2)-proces ved at løse et differensligningssystem. Generelt gælder

Sætning 9.19 Autokorrelationerne ρ_k for en AR(p)-proces $\phi(B)\tilde{Z}_t = A_t$ tilfredsstiller differensligningen

$$\phi(B)\rho_k = 0 \quad , \qquad k \neq 0$$

eller

$$\rho_k - \phi_1\rho_{k-1} - \cdots - \phi_p\rho_{k-p} = 0 \quad , \quad k \neq 0 \ .$$

Skrives disse ud for $k \leq p$, fås

$$\begin{aligned}
\rho_1 &= \phi_1 + \phi_2\rho_1 + \cdots + \phi_p\rho_{p-1} \\
\rho_2 &= \phi_1\rho_1 + \phi_2 + \cdots + \phi_p\rho_{p-2} \\
&\vdots \\
\rho_p &= \phi_1\rho_{p-1} + \phi_2\rho_{p-2} + \cdots + \phi_p \quad ,
\end{aligned}$$

de såkaldte Yule-Walker ligninger.

Bevis Ved at multiplicere begge sider af (13) med $\tilde{Z}_{t-k}$ fås

$$\tilde{Z}_{t-k}\tilde{Z}_t = \phi_1\tilde{Z}_{t-k}\tilde{Z}_{t-1} + \cdots + \phi_p\tilde{Z}_{t-k}\tilde{Z}_{t-p} + \tilde{Z}_{t-k}A_t \ ,$$

og ved at tage forventede værdier fås for $k > 0$

$$\gamma_k = \phi_1\gamma_{k-1} + \cdots + \phi_p\gamma_{k-p} \ ,$$

idet $E[\tilde{Z}_{t-k}A_t]$ da er 0. Ved division med γ_0 fås sætningen.

Q.E.D.

Sætning 9.20 Variansen for processen er

$$\sigma_z^2 = \frac{\sigma_a^2}{1-\rho_1\phi_1-\cdots-\phi_p\rho_p} \quad .$$

Bevis Med notation fra forrige bevis haves

$$\begin{aligned}\sigma_z^2 = \gamma_0 &= \phi_1\gamma_{-1} + \cdots + \phi_p\gamma_{-p} + E[\tilde{z}_t A_t] \\ &= \phi_1\gamma_1 + \cdots + \phi_p\gamma_p + \sigma_a^2 \quad .\end{aligned}$$

Q.E.D.

Bemærkning Den generelle løsning til differensligningen $\phi(B)\rho_k = 0$, hvor vi forudsætter, at rødderne i $\phi(z) = 0$ ligger uden for enhedscirklen, kan vises at være linearkombinationer af dæmpede sinus- og exponentialfunktioner, hvorfor autokorrelationsfunktionen for en stationer AR(p)-proces må være af denne form.

Eksempel 9.20 For en AR(1)-proces bliver Yule-Walker ligningerne med $\phi = \phi_1$

$$\begin{aligned}\rho_1 &= \phi \\ \rho_2 &= \phi\ \rho_1 \\ &\vdots \\ \rho_k &= \phi\ \rho_{k-1} \\ &\vdots\end{aligned}$$

o: vi har

$$\rho_k = \phi^k \quad , \quad k \geq 0 \quad .$$

(Bemærk, at stationaritet kræver $|\phi| < 1$).

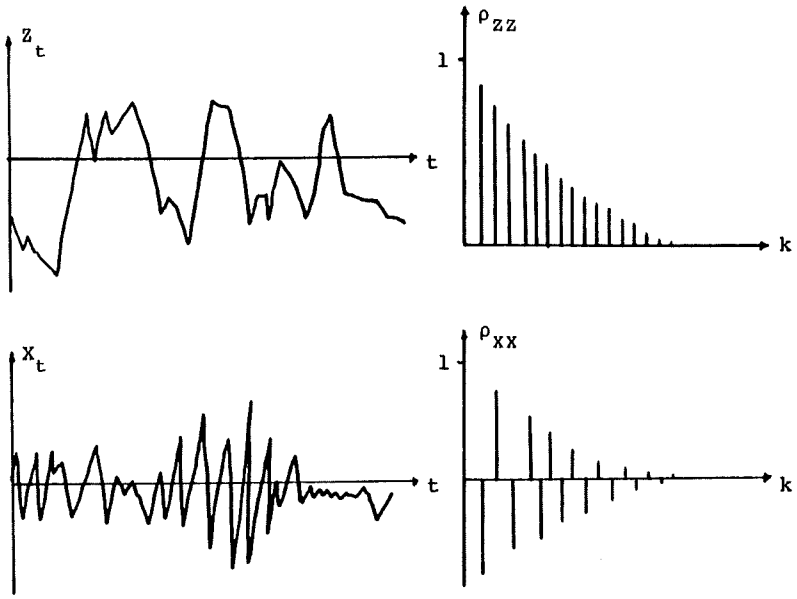
I nedenstående graf er vist realiserede udfald af to førsteordens autoregressive processer, nemlig

$$Z_t = 2 + 0.9\, Z_{t-1} + A_t$$

og

$$X_t = 2 - 0.9\, X_{t-1} + A_t \quad .$$

Endvidere er deres autokorrelationsfunktioner anført.



Realiserede udfald af AR(1)-processer og disses autokorrelationsfunktioner.

Vi ser, at processen med positivt ϕ_1 udviser langt færre svingninger end processen med negativt ϕ_1. Dette afspejles også i autokorrelationsfunktionerne, idet ρ_{ZZ} er positiv og ρ_{XX} har skiftende fortegn svarende til de observerede fluktuationer.

□

Ved analysen af stationære processer har vi hidtil i det væsentlige betjent os af autokorrelationer. Vi vil nu i analogi til den almindelige flerdimensionale analyse indføre en partiel autokorrelationsfunktion.

Definition 9.10 Lad $\tilde{Z}_t$ være en stationær, lineær proces. Vi definerer da den partielle autokorrelationskoefficient ϕ_{kk} ved

$$\phi_{kk} = \frac{\det \underline{\underline{U}}_k}{\det \underline{\underline{V}}_k} \quad ,$$

hvor

$$\underline{\underline{U}}_k = \sigma_z^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-2} & \rho_1 \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-3} & \rho_2 \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{k-4} & \rho_3 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & \rho_1 & \rho_k \end{bmatrix}$$

$$\underline{\underline{V}}_k = \sigma_z^2 \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-2} & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-3} & \rho_{k-2} \\ \rho_2 & \rho_1 & 1 & \cdots & \rho_{k-4} & \rho_{k-3} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & \rho_1 & 1 \end{bmatrix} .$$

Bemærkning 1 Vi bemærker, at $\underline{\underline{V}}_k = D((\tilde{Z}_1,\cdots,\tilde{Z}_1)')$, og at $\underline{\underline{U}}_k$ fremkommer af $\underline{\underline{V}}_k$ ved at erstatte sidste søjle med $(\rho_1,\cdots,\rho_k)'$.

Bemærkning 2 Hvis processerne er normale, kan det vises, at $\phi_{kk} = \mathrm{Cor}(X_1, X_{k+1} | X_2, \cdots, X_k)$.

Bemærkning 3 Af almindelige sætninger om løsningen af lineære ligningssystemer (Cramér's sætning p. 1.20) fremgår, at ϕ_{kk} kan fås ved løsning af Yule-Walker systemet.

$$\begin{bmatrix} 1 & \rho_1 & \cdots & \rho_{k-1} \\ \rho_1 & 1 & \cdots & \rho_{k-2} \\ \vdots & \vdots & & \vdots \\ \rho_{k-1} & \rho_{k-2} & & 1 \end{bmatrix} \begin{bmatrix} \phi_{k1} \\ \phi_{k2} \\ \vdots \\ \phi_{kk} \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_k \end{bmatrix} .$$

Vi har nu følgende to sætninger om partielle autokorrelationsfunktioner.

Sætning 9.21 For en AR(p)-proces gælder

$$\phi_{kk} \neq 0 \qquad k \leq p$$
$$\phi_{kk} = 0 \qquad k > p .$$

Bevis Forbigås, nærmest trivielt resultat.

Sætning 9.22 Den partielle autokorrelationsfunktion for en MA(q)-proces antager værdier forskellige fra 0 for vilkårligt store k . Den er domineret af dæmpede exponential- og/eller dæmpede sinusfunktioner.

Bevis Da processen $\tilde{Z}_t = \theta(B)A_t$ også kan skrives $\theta^{-1}(B)\tilde{Z}_t = A_t$, hvor θ^{-1} som den inverse til et polynomium har uendelig mange led, kan processen derfor opfattes som en uendelig autoregressiv proces, og resultatet forekommer ikke urimeligt ved sammenligning med sætning 9.21.

Bemærkning Vi ser, at der hersker en dualitet mellem stationaritetskrav og invertibilitetskrav for MA- og AR-processer og en tilsvarende dualitet mellem disse processers auto- og partielle autokorrelationsfunktioner.

En sammenfatning af AR- og MA-processerne er ARMA-processerne, de autoregressivt-glidende gennemsnitsprocesser.

En ARMA(p,q)-proces defineres i

Definition 9.11 $\tilde{Z}_t$ kaldes en ARMA(p,q)-proces, hvis den tilfredsstiller

$$\tilde{Z}_t = \phi_1\tilde{Z}_{t-1} + \phi_2\tilde{Z}_{t-2} + \cdots + \phi_p\tilde{Z}_{t-p} + A_t - \theta_1 A_{t-1} - \theta_2 A_{t-2} - \cdots - \theta_q A_{t-q} \quad (15)$$

d.v.s.

$$(1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p)\tilde{Z}_t =$$
$$(1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q)A_t ,$$

eller

$$\phi(B)\tilde{Z}_t = \theta(B)A_t , \quad (16)$$

hvor ϕ er et p'te grads og θ et q'te grads polynomium.

En ARMA(p,q)-proces kan enten opfattes som en autoregressiv proces med en MA-proces som residualproces, eller som en MA-proces med en AR-residualproces, i.e.

1) $\phi(B)\tilde{Z}_t = E_t$, $E_t = \theta(B)A_t$

2) $\tilde{Z}_t = \theta(B)C_t$, $\phi(B)C_t = A_t$.

Det er her åbenbart, at 1) netop beskriver en autoregressiv proces med et glidende gennemsnits residual. Ad 2) bemærker vi, at

$$\phi(B)\tilde{Z}_t = \phi(B)\theta(B)C_t = \theta(B)\phi(B)C_t = \theta(B)A_t ,$$

således at den glidende gennemsnits proces $\tilde{Z}_t$ med den autoregressive residualproces C_t netop tilfredsstiller (16).

Ud fra de tilsvarende sætninger for AR(p)- og MA(q)-processer forekommer følgende sætning nærmest indlysende

Sætning 9.23 En ARMA(p,q)-proces er stationær, hvis $\phi(z) = 0$ har alle rødder uden for enhedscirklen, og invertibel, hvis $\theta(z) = 0$ har alle rødder uden for enhedscirklen.

Bevis Forbigås.

Der gælder ikke pæne formler for momenterne, men vi har dog

Sætning 9.24 Autokovariansfunktionen i en ARMA(p,q)-proces tilfredsstiller følgende differensligning

$$\gamma_k = \phi_1\gamma_{k-1} + \cdots + \phi_p\gamma_{k-p} + \gamma_{za}(k) - \theta_1\gamma_{za}(k-1) - \cdots - \theta_q\gamma_{za}(k-q) ,$$

hvor

$$\gamma_{za}(k) = E[\tilde{Z}_{t-k}A_t] = \begin{cases} 0 , & k > 0 \\ \neq 0 , & k \leq 0 \end{cases} .$$

For variansen bliver ligningen

$$\gamma_0 = \phi_1\gamma_1 + \cdots + \phi_p\gamma_p + \sigma_a^2 - \theta_1\gamma_{za}(-1) - \cdots - \theta_q\gamma_{za}(-q) .$$

Bevis Forbigås.

Løsning af de i sætningen forekommende differensligninger giver i tilfældet p = q = 1

Corollar 1 I en ARMA(1,1)-model er

$$\gamma_0 = \frac{1 + \theta_1^2 - 2\phi_1\theta_1}{1 - \phi^2}\sigma_a^2$$

$$\gamma_1 = \frac{(1-\phi_1\theta_1)(\phi_1-\theta_1)}{1 - \phi_1^2}\sigma_a^2$$

$$\gamma_k = \phi_1\gamma_{k-1} , \qquad k \geq 2 .$$

Bevis Forbigås.

Ved betragtninger vedrørende generelle løsninger til ligninger af den i sætningen anførte form ledes man til

Corollar 2 Hvis $q < p$ i en ARMA(p,q)-proces, vil hele autokorrelationsfunktionen bestå af dæmpede eksponential- og/eller dæmpede sinusfunktioner. Hvis $q \geq p$, vil der være $q - p + 1$ begyndelsesværdier $\rho_0, \cdots, \rho_{q-p}$, som ikke følger dette generelle mønster.

Bevis Forbigås.

Endelig har vi

Sætning 9.25 Den partielle autokorrelationsfunktion for en ARMA(p,q)-proces antager værdier forskellige fra 0 for vilkårligt store k.

Bevis Processen $\phi(B)\tilde{Z}_t = \theta(B)A_t$ kan skrives $A_t = \theta^{-1}(B)\phi(B)\tilde{Z}_t$, d.v.s. som en uendeligordens autoregressiv proces. Heraf følger resultatet.

Q.E.D.

9.2.3 Den kontinuerte lineære proces

Den kontinuerte lineære proces defineres ganske analogt til den diskrete. Først må dog indføres begrebet "en kontinuert hvid støj" A_t.

Det volder lidt større problemer af matematisk art at definere A(t) end det tilsvarende diskrete begreb. Disse tekniske vanskeligheder omgår vi dog ved at definere en kontinuert "hvid støj"-proces som en proces med autokovariansfunktion

$$\gamma(u) = \sigma_A^2 \delta(u) ,$$

hvor δ er Dirac's deltafunktion. Processen har altså fuldstændigt ukorrelerede komponenter (da $\delta(u) = 0$, $u \neq 0$), men uendelig stor varians. Disse begreber kan begrundes ved grænseover-

gange i en proces, der har ukorrelerede tilvækster (Bachelier-Wiener-proces). Den interesserede læser henvises til litteraturen for nærmere detaljer. δ-funktionen er omtalt p. 9.9. Det skal indskærpes, at $V = \sigma_A^2 \neq V(A(t)) = \infty$.

Definition 9.12 Ved en kontinuert lineær proces $Z(t)$ forstås en proces af formen

$$Z(t) - \mu = \int_0^\infty \psi(u)A(t-u)du \ ,$$

hvor $A(t)$ er en kontinuert hvid støj. $Z(t) - \mu$ er output svarende til input $A(t)$ fra et lineært system med impulsresponsfunktion $\psi(u)$.

Vi giver momentfunktionerne for den kontinuert lineære proces i

Sætning 9.26 Vi betragter $Z(t)$ som defineret ovenfor. Da er

$$E(Z(t)) = \mu$$

$$V(Z(t)) = \sigma_A^2 \int_0^\infty \psi^2(u)du$$

$$Cov(Z(t),Z(t+u)) = \sigma_A^2 \int_0^\infty \psi(v)\psi(u+v)dv$$

Bevis Helt ligefremt. Vi har

$$E(Z(t)-\mu) = \int_0^\infty \psi(u)E(A(t-u))du = 0$$

Dette giver resultatet vedrørende middelværdien. Autokovariansen for $Z(t)$ er

$$\gamma_{ZZ}(u) = E([Z(t)-\mu][Z(t+u)-\mu])$$

$$= E\left[\int_0^\infty \psi(v)A(t-v)\,dv \int_0^\infty \psi(w)A(t+u-w)\,dw\right]$$

$$= \int_0^\infty \int_0^\infty \psi(v)\psi(w)E[A(t-v)A(t+u-w)]\,dvdw$$

$$= \int_0^\infty \int_0^\infty \psi(v)\psi(w)\gamma_{AA}(u+v-w)\,dvdw$$

$$= \int_0^\infty \int_0^\infty \psi(v)\psi(w)\sigma_A^2\delta(u+v-w)\,dvdw$$

$$= \sigma_A^2 \int_0^\infty \psi(v)\psi(u+v)\,dv \quad .$$

Q.E.D

Man kan nu i fuldstændig analogi med det tidligere definere MA, AR og ARMA-processer. Vi taler om en kontinuert MA-proces, såfremt $\psi(u)$ er 0 uden for et endeligt interval. Den kontinuerte AR-proces $Z(t)$ er givet som output efter input $A(t)$ fra et lineært system givet ved differentialligningen

$$z(t) = \phi_1 \frac{d\,z(t)}{dt} + \phi_2 \frac{d^2\,z(t)}{dt^2} + \cdots + \phi_p \frac{d^p\,z(t)}{dt^p} + a(t)$$

og den kontinuerte ARMA-proces som output fra et system givet ved ligningen

$$\phi_0\, z(t) - \phi_1 \frac{d\,z(t)}{dt} - \cdots - \phi_p \frac{d^p\,z(t)}{dt^p}$$

$$= \theta_0\, a(t) - \theta_1 \frac{d\,a(t)}{dt} - \cdots - \theta_q \frac{d^q\,a(t)}{dt^q} \quad .$$

Vi anfører en enkelt stationaritetsbetingelse i

Sætning 9.27 AR(p)-processen er stationær, såfremt rødderne i det karakteristiske polynomium

$$\phi(z) = 1 - \phi_1 z - \phi_2 z^2 - \cdots - \phi_p z^p = 0$$

har negative realdele.

Bevis Forbigås.

Bemærkning Jævnfør det tilsvarende resultat om stabilitet af et lineært system p. 9.23.

Vi begragter nu den kontinuerte AR(1)-proces i

Eksempel 9.21 Sætter vi p lig 1 i definitionsligningen for AR(1)-processen, fås

$$Z(t) - \mu = \phi_1 \frac{d\,Z(t)}{dt} + A(t) \quad .$$

Den karakteristiske ligning er

$$1 - \phi_1 z = 0$$

med roden

$$z = \frac{1}{\phi_1} \quad ,$$

så vi må kræve, at ϕ_1 er negativ, hvis vi ønsker stationaritet.

Ifølge sætning 9.9, p. 9.20, er frekvensresponsfunktionen for det til processen svarende lineære system

$$H(f) = \frac{1}{1-\phi_1 \cdot i2\pi f} = \frac{1}{1+|\phi_1| i2\pi f}$$

og af eksempel 9.2, p. 9.4, samt sætning 9.2, p. 9.13, følger, at impulsresponsfunktionen er

$$h(u) = \begin{cases} \frac{1}{|\phi_1|} e^{-u/|\phi_1|} & , \quad u \geq 0 \\ 0 & , \quad u < 0 \end{cases} .$$

Derfor er autokovariansfunktionen for Z(t) lig

$$\gamma(u) = \sigma_A^2 \int_0^\infty \frac{1}{\phi_1^2} e^{-v/|\phi_1|} e^{-(u+v)/|\phi_1|} dv$$

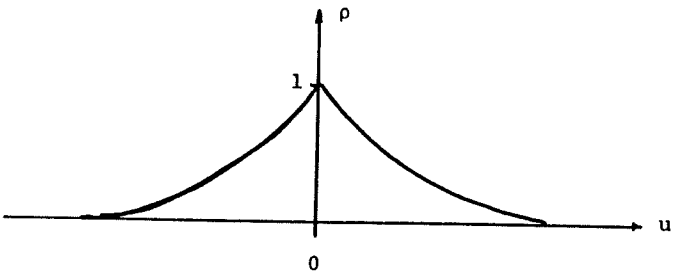
$$= \sigma_A^2 \frac{1}{\phi_1^2} e^{-u/|\phi_1|} \int_0^\infty e^{-2v/|\phi_1|} dv$$

$$= \sigma_A^2 \frac{1}{2|\phi_1|} e^{-u/|\phi_1|} \quad , \quad u > 0 .$$

Autokorrelationen bliver derfor

$$\rho(u) = e^{-|u|/|\phi_1|} \quad , \quad \forall u .$$

Den er afbildet i nedenstående graf



□

9.2.4 Estimation af autokovariansfunktion

Vi skal i det følgende ikke længere opretholde en strikte adskillelse mellem gennemgangen af resultater for diskrete og for kontinuerte processer. I de fleste tilfælde gælder der analoge

resultater, og vi vil da ofte kun eksplicit angive det, der gælder for f.eks. de kontinuerte processer.

Hvis man kun har en realisation x(t) givet i et interval [0,T] af en kontinuert proces X(t), kan det vises, at et rimeligt estimat for autokovariansfunktionen er

$$c_{xx}(u) = \frac{1}{T}\int_0^T (x(t)-\bar{x})\,(x(t+u)-\bar{x})\;dt$$

for $u \in [-T,T]$, hvor

$$\bar{x} = \frac{1}{T}\int_0^T x(t)\;dt\ .$$

Vi skal her ikke komme ind på det problem, hvornår man kan erstatte ensemble-gennemsnit med tidsgennemsnit (ergode-problemet); men blot godtage ovenstående estimater som værende intuitivt rimelige.

Ovenstående estimator kan under hensyntagen til, at x(t) er 0, uden for intervallet [0,T] skrives

$$c(u) = c_{XX}(u) = \begin{cases} \frac{1}{T}\int_0^{T-|u|} (X(t)-\bar{X})\,(X(t+|u|)-\bar{X})dt & 0\le|u|\le T \\ 0 & |u|>T\ , \end{cases}$$

hvor vi har indsat X(t) i stedet for x(t).

En alternativ estimator, der ofte anvendes, er

$$c'_{XX}(u) = \begin{cases} \frac{1}{T-|u|}\int_0^{T-|u|} (X(t)-\bar{X})\,(X(t+|u|)-\bar{X})\;dt & 0\le|u|\le T \\ 0 & |u|\le T \end{cases}$$

Det kan let vises, at

$$E[c_{XX}(u)] = \begin{cases} \gamma_{XX}(u)\,(1-\frac{|u|}{T}) & |u| \leq T \\ 0 & |u| > T \end{cases}$$

og

$$E[c'_{XX}(u)] = \begin{cases} \gamma_{XX}(u) & |u| \leq T \\ 0 & |u| > T \end{cases}$$

Med andre ord er $c'_{XX}(u)$ altså et centralt skøn over $\gamma_{XX}(u)$ for $|u| \leq T$, hvorimod $c_{XX}(u)$ kun er asymptotisk central for $T \to \infty$.

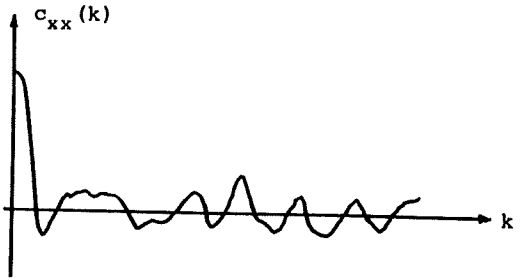
Hvis observationerne kommer fra en diskret tidsrække, d.v.s. der foreligger observationer $X_1,\ldots,X_N$, erstattes estimaterne af

$$c(k) = c_{XX}(k) = \frac{1}{N} \sum_{t=1}^{N-k} (X_t - \overline{X})\,(X_{t+k} - \overline{X})$$

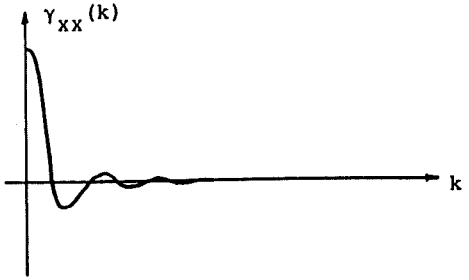
for $k = 0,1,\ldots,N-1$. Her er

$$\overline{X} = \frac{1}{N} \sum_{t=1}^{N} X_t \ .$$

Når man i praksis estimerer en autokovariansfunktion (eller en autokorrelationsfunktion ved at dividere autokovariansen med dens værdi i 0), vil man ofte få et billede i stil med følgende (efter Jenkins & Watts (1968))



d.v.s. en række svingninger, der ikke forsvinder. Dette skyldes ikke nødvendigvis, at observationer er relativt kraftigt korrelerede selv med meget store tidslag. F. eks. er ovenstående et estimat af en korrelationsfunktion beregnet på basis af 100 observationer fra en (simuleret) proces med følgende korrelationsfunktion



Forklaringen på dette fænomen er, at estimaterne af autovariansfunktionens værdier i sig selv er korrelerede. Det kan mere præcist vises, at

$$\mathrm{Cov}(c_{XX}(u_1),\ c_{XX}(u_2)) =$$

$$\frac{1}{T^2} \int_{-(T-u_1)}^{T-u_2} \varphi(r)\,[\gamma_{XX}(r)\gamma_{XX}(r+u_2-u_1)+\gamma_{XX}(r+u_2)\gamma_{XX}(r-u_1)]\,dr\ ,$$

hvor

$$\varphi(r) = \begin{cases} T-u_2-r\ , & r \geq 0 \\ T-u_2\ , & -(u_2-u_1) \leq r \leq 0 \\ T-u_1+r\ , & -(T-u_1) \leq r \leq -(u_2-u_1)\ . \end{cases}$$

For rimeligt store værdier af T kan det vises, at

$$\mathrm{Cor}(c_{XX}(u_1),\ c_{XX}(u_2)) \simeq$$

$$\frac{1}{T} \int_{-\infty}^{\infty} [\gamma_{XX}(r)\gamma_{XX}(r+u_2-u_1)+\gamma_{XX}(r+u_2)\gamma_{XX}(r-u_1)]\,dr\ .$$

For diskrete data er den tilsvarende approximation

$$\mathrm{Cor}(c_{xx}(k),\ c_{xx}(\ell)) \simeq$$

$$\frac{1}{N} \sum_{r=-\infty}^{\infty} [\gamma_{xx}(r)\gamma_{xx}(r+\ell-k) + \gamma_{xx}(r+\ell)\gamma_{xx}(r-k)]$$

En nærmere diskussion af disse vigtige formler kan eksempelvis findes i Jenkins & Watts (1968).

Hvis stationaritetskravene ikke er tilfredsstillede, kan man ofte "fjerne" ikke-stationariteten ved at foretage en passende filtrering af ens data (om filtre: se afsnit 9.5). I en række standardprogrammer (e.g. BMD02T) findes der en mulighed for at anvende et filter defineret ved

$$y_t = x_t - ax_{t-1} ,$$

hvor a $\in[-1,1]$. Det kan da vises (se e.g. Jenkins & Watts (1968)), at

$$c_{yy}(k) \simeq -ac_{xx}(k-1) + (1+a^2)\, c_{xx}(k) - ac_{xx}(k+1),$$

således at man rimeligt nemt kan komme fra x-rækkens autokovarians til y-rækkens.

9.3 Den klassiske analyse

Vi anfører i dette delafsnit kort nogle af de vigtigste elementer i den klassiske analyse af tidsrækker. Dette sker uden den mindste brug af teori, og en række af de elementer, man tidligere anvendte, omtales ikke, idet de nu har fundet bedre løsninger ved behandling af rækkerne i frekvensdomænet, jvf. afsnit 9.4.

9.3.1 Udjævning og trend

Ofte udviser en tidsrække nogle relativt hurtige svingninger, som i mange sammenhænge a priori skønnes at være uden væsentlig

interesse. Der har derfor igennem mange år været stor interesse for at udvikle metoder, der kan "fjerne" disse svingninger. I dette afsnit skal vi betragte to klassiske former for udjævningsmetoder, men i et senere afsnit om filtrering skal vi give en mere systematisk fremstilling af emnet.

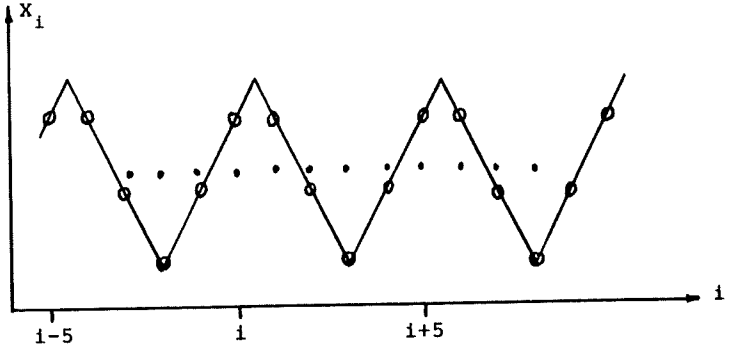
Først betragter vi den såkaldte glidende gennemsnitsmetode. Lad observationerne være

$$\ldots, x_i, \ldots, x_{i+n}, \ldots \quad ,$$

og lad os antage, at fluktuationerne synes at være af perioden 2p+1, som skitseret nedenfor (i et noget idealiseret tilfælde med 2p+1 = 5). x angiver de enkelte målinger (i, x_i). Vi beregner nu størrelsen

$$z_i = \frac{x_{i-p} + x_{i-p+1} + \ldots + x_i + x_{i+1} + \ldots + x_{i+p}}{2p+1}$$

$$= \frac{1}{2p+1} \sum_{j=-p}^{p} x_{i+j} \quad .$$



o observation (i, x_i)

• glidende gennemsnit af orden 5.

Disse værdier z_i kaldes glidende gennemsnit af orden 2p+1 af x_i'erne. Den enkelte værdi x_i er erstattet af gennemsnittet af x_i og dens 2p nabopunkter. De er - stadig med 2p+1 = 5 - angivet som · på figuren, og det ses, at z_i'erne er konstante, d.v.s. det anførte glidende gennemsnit har fuldstændigt elimineret svingningen med perioden 2p+1 = 5. Det ses let, at f. eks. et glidende gennemsnit af orden 2p-1 ikke har denne udjævnende effekt.

Hvis perioden er af størrelsen 2p, er det ikke helt så åbenbart, hvad man skal anføre som glidende gennemsnit. Det er let nok successivt at beregne gennemsnit af 2p på hinanden følgende værdier, men der er så intet naturligt centerpunkt, hvor disse værdier kan anbringes. I stedet beregnes

$$u_i = \frac{1}{2p} \sum_{j=-p}^{p-1} x_{i+j}$$

$$v_i = \frac{1}{2p} \sum_{j=-p+1}^{p} x_{i+j} ,$$

og vi sætter da den udjævnede værdi lig

$$z_i = \frac{u_i+v_i}{2} = \frac{1}{4p} (x_{i-p}+2x_{i-p+1}+\ldots+2x_i+\ldots+2x_{i+p-1}+x_{i+p}) .$$

Et resultat med forbindelse til det introducerende eksempel er følgende sætning om udglatning af en tidsrække med en underliggende cosinussvingning.

Sætning 9.28 Vi betragter en stokastisk proces $\ldots,X_1,X_2,\ldots$, med

$$E(X_t) = \cos(2\pi ft-\theta) ,$$

d.v.s. en cosinussvingning med periode $1/f$. Sættes z_i lig det glidende gennemsnit af orden $2p+1$, d.v.s.

$$Z_t = \frac{1}{2p+1}(X_{t-p} + \cdots + X_{t+p})$$

er

$$E(Z_t) = \frac{\sin(\pi f(2p+1))}{(2p+1)\sin(\pi f)} \cos(2\pi ft-\theta) \quad .$$

Hvis specielt $2p+1 = 1/f$ (d.v.s. lig perioden), er

$$E(Z_t) = 0 \quad .$$

Bevis Sætningen følger ved at skrive $\cos x = \mathrm{re}(\exp(ix))$ og benytte formlen for en kvotientrække. Detaljerne forbigås.

Bemærkning Vi ser, at effekten af udjævningen ved hjælp af et glidende gennemsnit er at reducere størrelsen af middelværdifunktionen. For en lang periode, i.e. en lille værdi af f, vil $\sin(\pi f(2p+1))/\sin(\pi f) \simeq 2p+1$, hvorfor $E(Z_t) \simeq E(X_t)$, d.v.s. der sker ingen væsentlig reduktion. For fastholdt periode vil $|E(Z_t)|$ aftage for voksende p (med $2p+1 < 1/f$) og altså forsvinde helt for $2p+1 = 1/f$.

Man kan selvfølgelig også anvende glidende gennemsnit med forskellige vægte på de enkelte observationer. En meget anvendt klasse af sådanne fremkommer ved som udglattet værdi y_i at anvende et mindste kvadraters skøn (baseret på målingerne $(x_{t-p},\cdots,x_{t+p})$) af et polynomium. Vi anfører resultatet for et anden og tredje grads polynomium i nedenstående sætning. I Kendall & Stuart (1966) gives en række andre resultater.

Sætning 9.29 Lad der være givet en tidsrække

$$\cdots,x_{t-p},\cdots,x_t,\cdots,x_{t+p},\cdots \quad .$$

Da antager det tredie (og anden) grads polynomium, man får ved at "fitte" punkterne $(t-2,x_{t-2}),\cdots,(t+2,x_{t+2})$ ved hjælp af mindste kvadraters metode, værdien

$$Y_t = \frac{1}{35}(-3x_{t-2} + 12x_{t-1} + 17x_t + 12x_{t+1} - 3x_{t+2})$$

i punktet t .

Bevis Vi arbejder med følgende model

$$E(X_t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3 \ ,$$

og vi skal estimere (a_0, a_1, a_2, a_3) på basis af værdierne

$$(t-2, X_{t-2}),\ (t-1, X_{t-1}),\ (t, X_t),\ (t+1, X_{t+1}),\ (t+2, X_{t+2}) \ .$$

Vi flytter tidsaksens nulpunkt og betragter punkterne

$$(-2, X_{-2}),\ (-1, X_{-1}),\ (0, X_0),\ (1, X_1),\ (2, X_2) \ .$$

Vi skal minimalisere

$$\varphi = \sum_t (X_t - a_0 - a_1 t - a_2 t^2 - a_3 t^3)^2 \ ,$$

hvor t antager værdierne -2, -1, 0, 1, 2 . Partiel differentiation giver

$$\frac{\partial \varphi}{\partial a_0} = -2 \sum (X_t - a_0 - a_1 t - a_2 t^2 - a_3 t^3)$$

$$\frac{\partial \varphi}{\partial a_1} = -2 \sum (tX_t - a_0 t - a_1 t^2 - a_2 t^3 - a_3 t^4)$$

$$\frac{\partial \varphi}{\partial a_2} = -2 \sum (t^2 X_t - a_0 t^2 - a_1 t^3 - a_2 t^4 - a_3 t^5)$$

$$\frac{\partial \varphi}{\partial a_3} = -2 \sum (t^3 X_t - a_0 t^3 - a_1 t^4 - a_2 t^5 - a_3 t^6) \ .$$

Vi bemærker nu, at $\sum t^p = 0$, hvis p er ulige, og t antager de anførte værdier. De relevante summer af lige potenser bliver

$$\sum t^2 = 10$$

$$\sum t^4 = 34$$

$$\sum t^6 = 130$$

sættes de partielle afledede lig 0 , og udnyttes ovenstående, fås ligningssystemet

$$\sum X_t = 5a_0 + 10a_2$$

$$\sum t X_t = 10a_1 + 34a_3$$

$$\sum t^2 X_t = 10a_0 + 34a_2$$

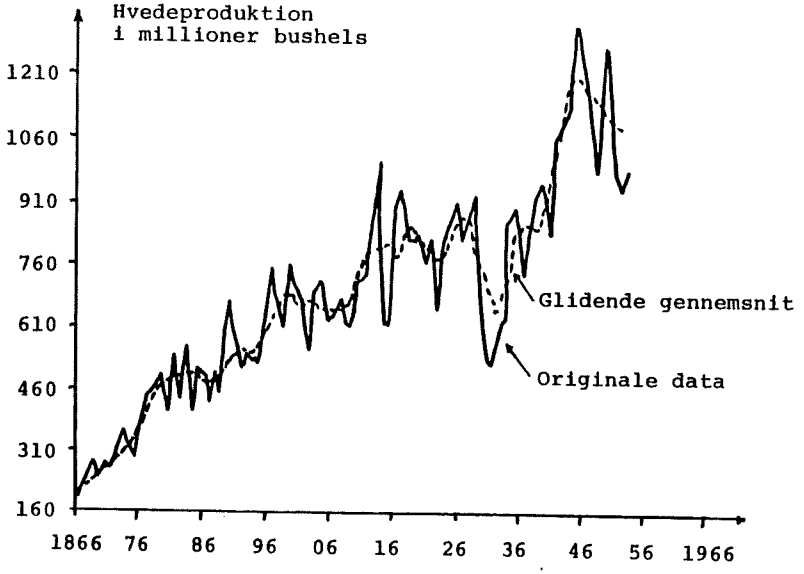
$$\sum t^3 X_t = 34a_1 + 130a_3 \ .$$

Vi skal alene bestemme a_0 og kan derfor nøjes med at betragte den første og den tredje ligning. Løsningen af disse giver det i sætningen anførte resultat. Ved at gennemføre de tilsvarende betragtninger for et andengrads polynomium ses, at formlen bliver den samme.

Q.E.D.

Vi skal nu vise, hvorledes et almindeligt glidende gennemsnit anvendt på et større talmateriale tager sig ud.

Eksempel 9.22 Nedenstående graf er baseret på 100 års hvedeproduktion i USA (årene 1866-1966), jfr. Chakravarti, Laha & Roy (1967). I figuren vises dels de oprindelige observationer og dels et 5-års glidende gennemsnit. Disse ses at have væsentligt mindre udsving end den oprindelige række.



Hvedeproduktion i USA, 1866-1966.

□

En anden hyppigt anvendt udglatningsmetode er eksponentiel udjævning. Vi betragter igen en række observationer

$$x_1, \ldots, x_{i+n}, \ldots \quad .$$

Vi definerer de udjævnede observationer y_i ved

$$\text{(*)} \qquad y_i = ax_i + (1-a)y_{i-1} \quad , \quad 0 < a < 1 \ ,$$

d.v.s. den aktuelle y_i sættes lig et vejet gennemsnit af den observerede værdi x_i og den foregående værdi af de udjævnede observationer. Vi siger, at observationerne y_i er fremkommet ved en første ordens eksponentiel udjævning.

Højere ordens former for eksponentiel udjævning defineres på helt tilsvarende måde (jfr. afsnittet om autoregressive processer af dels første og dels højere orden i det foregående). Ved rekursiv løsning af (*) fås

$$y_i = a \sum_{\nu=0}^{i-1} (1-a)^{\nu} x_{i-\nu} + (1-a)^{i} y_0 \ ,$$

hvor værdien y_0 er en mere eller mindre arbitrært valgt begyndelsesværdi. Hvis y_0 sættes lig x_1, bliver y_1 lig x_1. Af ovenstående indses, at en observation taget for ν tidsenheder siden indgår med vægten

$$a(1-a)^{\nu} \quad , \quad 0 \leq a \leq 1 \ .$$

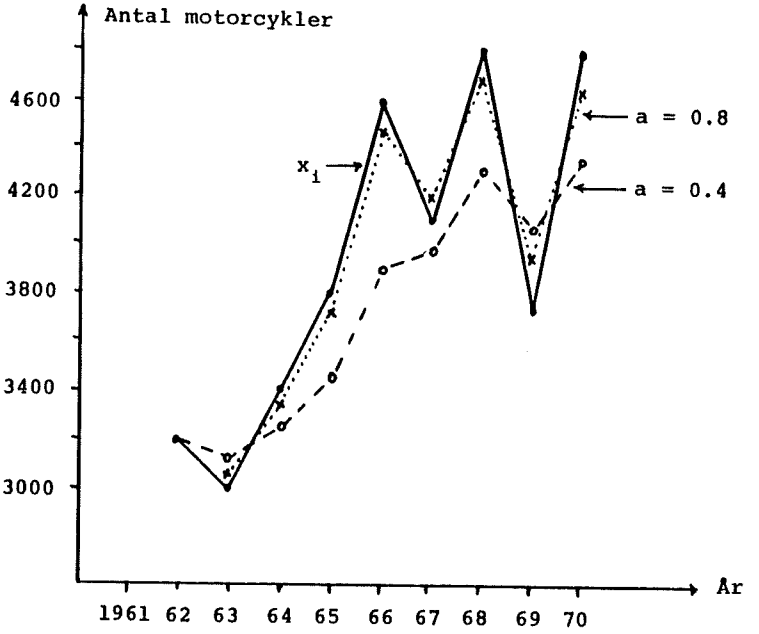
Det ses, at vægtene er eksponentielt (geometrisk) aftagende, deraf navnet eksponentiel udjævning. Hvis a er meget nær ved 1, lægges der forholdsmæssigt mere vægt på de nyeste observationer, end hvis a er nær 0. Ved at vælge forskellige værdier af a er det derfor muligt at justere udjævningens "hukommelse". Vi skal illustrere problemerne ved et lille eksempel.

Eksempel 9.23 I nedenstående tabel er anført det antal motorcykler, der er produceret i årene 1962-70 hos B.K. Thornton Company. Data er fra Clelland, deCani & Brown (1973). Endvidere er anført eksponentielt udglattede værdier med a = 0.4 og a = 0.8.

År	i	Antal prod. m.c. x_i	Eksp. udgl. a = 0.4	Eksp. udgl. a = 0.8
1961	0		3200	3200
1962	1	3200	3200	3200
1963	2	3000	3120	3040
1964	3	3400	3232	3328
1965	4	3800	3459	3706
1966	5	4600	3915	4421
1967	6	4100	3989	4164
1968	7	4800	4313	4673
1969	8	3700	4067	3895
1970	9	4800	4360	4619

I nedenstående figur er såvel de oprindelige som de udglattede data anført. Det ses, at data svarende til a = 0.4 udviser et

langt mere "glat" forløb end data svarende til 0.8. Disse sidst udglattede "følger" i langt højere grad de oprindelige observationer - dog stadig med mindre udsving end disse.



Eksponentielt udglattede motorcykelproduktionsdata.

□

Hvis man vil prediktere kommende værdier ved hjælp af eksponentiel udjævning, kan man eventuelt bestemme en "optimal" værdi af a ved hjælp af mindste kvadraters metode, i.e. ved at minimalisere

$$\sum_{t=-\infty}^{t_0}\left[x_t - a\sum_{\nu=1}^{k}(1-a)^{\nu}x_{t-\nu}\right]^2 \simeq \sum_{t=-\infty}^{t_0}\left(x_t - y_{t-1}^{(a)}\right)^2$$

med hensyn til a. Her er k valgt så stor, at $(1-a)^k$ i denne forbindelse kan anses for forsvindende. Tilsvarende antager t i praksis selvfølgelig kun endeligt mange værdier. $y_{t-1}^{(a)}$ er den med vægten a exponentielt udglattede værdi til tid t-1.

Som den forudsagte værdi til tid t vælges så $\hat{x}_t$ defineret ved

$$\hat{x}_t = y_{t-1} = ax_{t-1} + (1-a)\, y_{t-2}$$
$$= ax_{t-1} + (1-a)\, \hat{x}_{t-1} .$$

Når tid t-1 er indtruffet, kan vi forudsige den kommende værdi til tid t ved at opdatere den gamle forudsigelse $\hat{x}_{t-1}$. Denne opdatering sker ved en simpel vejet gennemsnitsdannelse mellem $\hat{x}_{t-1}$ og den faktisk indtrufne observation x_{t-1}.

Denne metodes adaptive karakter bliver mere fremtrædende, når vi vælger følgende omskrivning

$$\hat{x}_t = \hat{x}_{t-1} + a(x_{t-1} - \hat{x}_{t-1})$$
$$= \hat{x}_{t-1} + ae_t ,$$

hvor e_t er den observerede forudsigelsesfejl.

I Brown (1963) er bl.a. behandlet estimationsproblematikken vedrørende a i forskellige specialtilfælde. Sammenlign også med afsnittet om Box-Jenkins metode.

Et problem, der er beslægtet med udjævning af observationsrækker, er bestemmelse af en såkaldt trend eller tendens, om man vil (jfr. bemærkningen om trend i næste afsnit).

Vi vil her først søge at approksimere trenden med et polynomium af passende grad.

Den ene metode er at bruge orthogonale polynomier, som det er beskrevet i afsnit 4.2. Man bestemmer da successivt polynomier af højere og højere grad, indtil en forøgelse af graden ikke formindsker residualkvadratafvigelsessummen væsentligt mere.

Den anden metode går ud fra, at observationerne X_t kan skrives

$$X_t = p_k(t) + A_t ,$$

hvor A_t er et stokastisk (stationært) bidrag, og $p_k(t)$ er et polynomium eller en funktion, der tilstrækkelig godt kan beskrives ved et polynomium i t af k'te grad.

Vi finder

$$\begin{aligned}\nabla X_t &= X_t - X_{t-1} \\ &= p_k(t) - p_k(t-1) + \nabla A_t \ .\end{aligned}$$

Det ses umiddelbart, at $p_k(t) - p_k(t-1)$ er et polynomium af k-1'te grad. Ved at tage k+1 successive differenser er det derfor muligt fuldstændigt at eliminere den polynomiale trend, idet

$$\nabla^{k+1} X_t = \nabla^{k+1} p_k(t) + \nabla^{k+1} A_t = \nabla^{k+1} A_t$$

Problemet er nu at bestemme k. Vi sætter

$$V_j = \frac{\sum_t (\nabla^j X_t)^2}{(N-j)\binom{2j}{j}}$$

Hvis x_t'erne er ukorrelerede med samme varians σ^2, da gælder

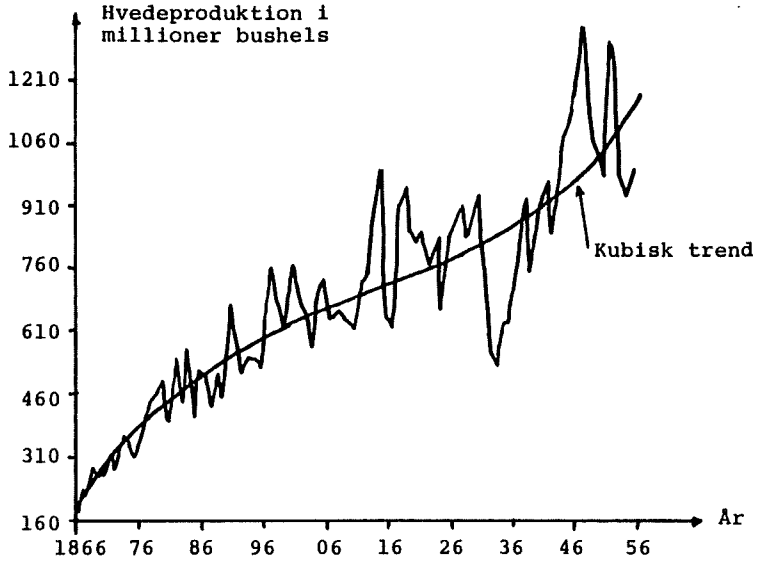
$$E(V_j) = \begin{cases} \delta_j^2 > \sigma^2 & j \leq k \\ \sigma^2 & j > k \end{cases}$$

(se f. eks. Kendall (1973)). Man kan derfor bestemme den nødvendige grad af et polynomium ved at beregne størrelserne $v_1, v_2, \ldots$ etc. og dernæst bemærke, fra hvilket trin disse synes ens (bortset fra tilfældige fluktuationer). Metoden kaldes "the variate difference method" i den angelsaxiske litteratur.

Eksempel 9.24 Vi betragter de i eksempel 22 citerede data over hvedeproduktionen i USA. For disse tal gælder

i :	1	2	3	4	5	6
V_i :	6906.38	5368.43	4645.17	4290.10	4098.73	3997.69

Det ses, at V_4, V_5 og V_6 er af nogenlunde samme størrelsesorden, og vi vil derfor approksimere trenden med et polynomium af tredie grad. Dette bestemmes ved en almindelig regressionsanalyse, og polynomiet er indtegnet i nedenstående figur.



Hvedeproduktion i USA, 1866-1956, med kubisk trend indtegnet.

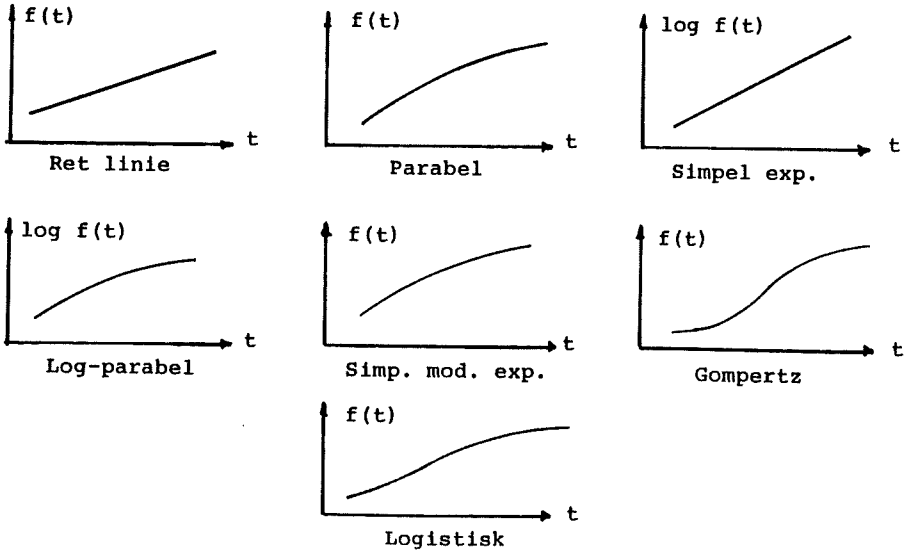
Der ses at være en smuk overensstemmelse mellem den "visuelt" bestemte tendens og den kubiske trend.

□

Der findes selvsagt en mængde andre måder at bestemme trendkurver på. En systematisk fremstilling er givet i Gregg, Hossell & Richardson (1964). Man opererer her med 7 familier af trendkurver. De er angivet i skemaet p. 9.79, og deres grafer er skitseret p. 9.79.

Ved bestemmelse af den relevante kurve anvendes en række transformationer, hvor hovedideen er at bestemme en relation mellem

Trendkurve	Ligning	Differentialrelation
Ret linie	$a+bt$	$f'(t) = b$
Parabel	$a+bt+ct^2$	$f'(t) = b+2ct$
Simpel eksponentiel	$\exp(a+bt)$	$\frac{f'(t)}{f(t)} = b$
Logaritmisk parabel	$\exp(a+bt+ct^2)$	$\frac{f'(t)}{f(t)} = b+2ct$
Simpel modificeret eksponentiel	$a-br^t$, $0 \leq r \leq 1$	$\log f'(t) = \log(-b \log r)+(\log r)t$
Gompertz	$\exp(a-br^t)$	$\log \frac{f'(t)}{f(t)} = \log(-b \log r)+(\log r)t$
Logistisk	$1/(a+br^t)$, $0<r<1$	$\log \frac{f'(t)}{f(t)^2} = \log(-b \log r)+(\log r)t$



Trendkurver.

f og f', som er lineær i tiden. Disse relationer er ligeledes anført i skemaet under betegnelsen differentialrelation.

Identifikationen, d.v.s. fastlæggelse af den relevante familie, foregår nu mere præcist, ved at man "estimerer" f(t) ved hjælp af et glidende gennemsnit af orden 2p+1, hvor p sædvanligt (men mere eller mindre arbitrært) sættes til 2-5, i.e.

$$\hat{f}(t) = \frac{1}{2p+1} (y_{t-p} + \ldots + y_{t-1} + y_t + y_{t+1} + \ldots + y_{t+p}) .$$

Tilsvarende "estimeres" f'(t) ved

$$\hat{f}'(t) = \left[2 \sum_{j=1}^{p} j^2 \right]^{-1} \left[2p \frac{2y_{t+p} - y_{t-p}}{2p} + \ldots + 2 \frac{y_{t+1} - y_{t-1}}{2} \right]$$

$$= \frac{3}{p(p+1)(2p+1)} [-py_{t-p} - \ldots - y_{t-1} + y_{t+1} + \ldots + py_{t+p}] ,$$

der altså er et vejet gennemsnit af de p hældningskoefficienter, der er bestemt ved hjælp af målinger, der ligger symmetrisk omkring y_t.

Ved at plotte $\hat{f}(t)$ og $\hat{f}'(t)$, deres logaritmer samt de forskellige differentialrelationer kan man finde den bedst beskrivende trendkurve.

Når man har bestemt den relevante kurve, kan man fastlægge initiale estimater for parametrene ved hjælp af graferne; men mere tilfredsstillende skøn opnås selvfølgelig ved hjælp af f. eks. mindste kvadraters metode som anført i kapitel 4.

I Gregg, Howell & Richardson (1964) er givet en række nyttige regneskemaer, rekursive relationer m.v., ligesom der er anført kurveblade til hjælp ved fastlæggelse af konfidensintervaller.

9.3.2 Den klassiske dekomposition

I den "klassiske" tidsrækkeanalyse, som den især er blevet anvendt i forbindelse med økonomiske data, opererer man ofte med følgende dekomposition af den betragtede tidsrække

$$X_i = T_i + S_i + C_i + R_i ,$$

hvor

T_i angiver trend

S_i - sæsonvariation

C_i - cyklisk svingning af længere varighed

R_i - fejl eller residual.

Ofte anvendes også en rent multiplikativ model, i.e.

$$X_i = T_i \, S_i \, C_i \, R_i ,$$

eller blandinger af en additiv og en multiplikativ model. I det følgende vil vi dog koncentrere os om den additive model. Modifikationerne til de øvrige er forholdsvis åbenbare.

Vi vil først give en verbal skildring af de 4 involverede begreber. Vi følger her en meget "amerikansk" fremstilling (Clelland, deCani & Brown (1973)).

Trenden i en tidsrække er en identificerbar bevægelse, som sker over et langt tidsrum, sædvanligvis 25 år eller mere. To faktorer er særlig vigtige ved deres påvirkning af trenden i de fleste økonomiske tidsrækker, nemlig befolkningsændringer og tekniske fremskridt.

Klima og sæder er de to dominerende faktorer i sæsonvariation. Arbejdsløshedstallene for jord- og betonarbejdere stiger "altid" om vinteren, og legetøjsbranchens salg er "altid" stort i julemåneden.

Cykliske fluktuationer er en periodisk bevægelse med en periode (fra en spids til den næste) af fra to år til så meget som 15 eller 20. De kræfter, som inducerer cykliske bevægelser, er en hovedinteresse for dem, der er interesserede i økonomiske betingelser, men der hersker ingen enstemmighed med hensyn til troen på deres årsager. De er blevet tillagt faktorer så for-

skellige som solpletter eller massepsykologiske svingninger mellem optimisme og pessimisme.

Residualet er, groft sagt, de uregelmæssigheder, der ikke forklares ved de øvrige faktorer. Disse uregelmæssigheder skyldes tilfældighed eller uforudsigelige faktorer som flodbølger, krige, pest og lignende
- for stadig at holde os til ovenstående værk.

Problemet i analysen af et sådant system er at dekomponere tidsrækken i de 4 komponenter. Vi skal ikke komme ind på en mere teoretisk præget diskussion, men alene illustrere fremgangsmåden i et eksempel efter at have postuleret en metode.

Vi har altså modellen

$$X_i = T_i + S_i + C_i + R_i \; ; \quad i=1,\ldots,n \; .$$

Proceduren i dekompositionen er følgende. Vi fjerner først sæsonfaktoren S_i. Dette gøres ved udvikling af sæsonindices a_j, $j=1,\ldots,m$, hvor m angiver antallet af forskellige sæsonperioder. Vi skal vende tilbage til beregningen af disse sæsonindices i det følgende. Vi antager, at en evt. sæsonpåvirkning vil være af multiplikativ karakter, som man ofte ser det e.g. i økonomi o.l. (ændringer i omsætning fra julekvartal til sommerkvartal sker i % af omsætningen etc.). Følgelig justeres rækken for sæsonsvingninger ved dannelse af

$$X_i' = \frac{X_i}{a_j} = T_i + C_i + R_i$$

for en observation X_i, der falder i den j'te sæsonperiode. Vi finder da sæsoneffekten S_i som

$$S_i = X_i - X_i'$$

Dernæst fittes en trend T_i til rækken X_i' (f. eks. ved hjælp af regressionsanalyse). Dette giver anledning til rækken

$$X_i'' = X_i' - T_i = C_i + R_i \; ,$$

som nu er justeret for såvel trend som sæsonsvingning. Den cykliske komponent findes nu af $X_i'' = C_i + R_i$ ved at danne et glidende gennemsnit af samme orden som den formodede cykliske periode. Dette glidende gennemsnit vil eliminere komponenten C_i i rækken $C_i + R_i$, men også samtidig "forvride" residualet R_i. Antages det imidlertid, at "væsentlige" dele af R_i lades uberørt, kan C_i findes ved at subtrahere det fremkomne glidende gennemsnit fra X_i''. Det må dog præciseres, at dette sidste er en noget tvivlsom fremgangsmåde. Det vil være langt mere hensigtsmæssigt her at benytte et af de i afsnit 9.5 beskrevne båndpasfiltre ved isoleringen af den cykliske komponent.

Et andet og væsentlige forbehold over for den valgte metode er, at et glidende gennemsnit kun vil eliminere den betragtede cykliske svingning, hvis der er tale om en ren cosinussvingning med netop den angivne periode.

Vi vender os nu mod beregningen af sæsonindices. Vi danner det glidende gennemsnit af orden m af de oprindelige observationer X_i. Dette giver en række Z_i, som i det væsentlige er befriet for voldsomme sæsonsvingninger. Ud fra betragtninger analogt til de tidligere givne dannes dernæst rækken

$$\frac{X_i}{Z_i},$$

idet denne vil repræsentere målingen på den i'te dag relativt til "niveauet" på det pågældende tidspunkt. Vi danner derefter nogle foreløbige indices a_j' ved

$$a_j' = \left\{ \sum_{\substack{\text{alle "i" som} \\ \text{falder i en} \\ \text{sæsonperiode j}}} \frac{X_i}{Z_i} \right\} \Big/ (\text{antal led i sum}) \quad .$$

Disse tal $a_1', \ldots, a_m'$ normeres nu, således at de får en gennemsnitsværdi på 1, d.v.s. vi definerer det j'te sæsonindex a_j ved

$$a_j = \frac{a'_j \cdot m}{\sum_j a'_j} .$$

Vi anfører dernæst et illustrativt eksempel.

Eksempel 9.25 I tabellen næste side er anført 8 ugers observationer af døgngennemsnittet af SO_2-indholdet i luften i en dansk provinsby. Målingerne stammer fra eftersommeren 1973 (mandag den 30/7 og 8 uger frem). Måleenheden er µg SO_2/m^3. Vi vil betragte en opspaltning

$$X_i = T_i + S_i + R_i$$

trend ugeeffekt res.

I tabellen er også anført de udregninger, der er nødvendige for dekompositionen.

I figuren p. 9.87 er vist de grafiske billeder for de beregnede effekter. Vi præciserer her, at det er den nederste kurve R_i, der har interesse ud fra f. eks. et kontrolsynspunkt. Den angiver måleværdien på en enkelt dag, justeret for det sædvanlige forureningsmønster, der er på denne ugedag (e.g. vil målinger på 30 µg/m^3 på en søndag og en torsdag blive vurderet forskelligt, fordi vi oftest har større værdier midt i ugen.)

De i omstående tabel anførte værdier af a_j fremkommer ved at tage gennemsnit for de enkelte ugedage af værdierne af X_i/Z_i (for torsdags vedkommende f. eks. de 7 sidste torsdage. For de øvrige ugedage kan X_i/Z_i kun bestemmes 7 gange). Værdierne er anført p. 9.86.

Obs. nr.	Dag	Observation X_i	Glidende gennemsnit af orden 7 Z_i	X_i/Z_i	Obs. justeret for ugevariation $X_i/a_j = T_i+R_i$	Trend T_i	Obs. justeret for ugevar. og trend $X_i/a_j-T_i = R_i$	Ugevariation $S_i = X_i-X_i/a_j$
1	M	23	-	-	24.6	21.76	2.9	-1.6
2	T	9	-	-	7.2	21.74	-14.6	1.8
3	O	23	-	-	20.4	21.72	-1.3	2.6
4	T	32	22.0	1.46	22.9	21.69	1.2	9.1
5	F	52	21.7	2.40	44.1	21.67	22.5	7.9
6	L	9	25.9	0.35	14.9	21.65	-6.7	-5.9
7	S	6	24.7	0.24	11.8	21.62	-9.8	-5.8
8	M	21	27.3	0.77	22.5	21.60	0.9	-1.5
9	T	38	22.4	1.69	30.3	21.58	8.7	7.7
10	O	15	22.4	0.67	13.3	21.55	-8.3	1.7
11	T	50	22.9	2.19	35.8	21.53	14.3	14.2
12	F	18	22.1	0.81	15.3	21.51	-6.2	2.7
13	L	9	25.7	0.35	14.9	21.48	-6.6	-5.9
14	S	9	31.6	0.29	17.8	21.46	-3.7	-8.8
15	M	16	30.7	0.52	17.1	21.44	-4.3	-1.1
16	T	63	35.7	1.76	50.3	21.41	28.9	12.7
17	O	56	36.3	1.54	49.6	21.39	28.2	6.4
18	T	44	35.9	1.23	31.5	21.37	10.2	12.5
19	F	53	35.7	1.48	45.0	21.34	23.6	8.0
20	L	13	28.4	0.46	21.5	21.32	0.2	-8.5
21	S	6	22.6	0.27	11.8	21.30	-9.5	-5.8
22	M	15	18.4	0.81	16.1	21.20	-5.2	-1.1
23	T	12	13.0	0.92	9.6	21.25	-11.7	2.4
24	O	15	15.0	1.00	13.3	21.22	-7.9	1.7
25	T	15	17.6	0.85	10.7	21.20	-10.5	4.3
26	F	15	19.7	0.76	12.7	21.18	-8.4	2.3
27	L	27	20.6	1.31	44.7	21.15	23.6	-17.7
28	S	24	20.6	1.17	47.4	21.13	26.2	-23.4
29	M	30	20.6	1.46	32.1	21.11	11.0	-2.1
30	T	18	20.6	0.88	14.4	21.08	-6.7	3.6
31	O	15	18.0	0.83	13.3	21.06	-7.8	1.7
32	T	15	15.4	0.97	10.7	21.04	-10.3	4.3
33	F	15	13.3	1.13	12.7	21.01	-8.3	2.3
34	L	9	14.1	0.64	14.9	20.99	-6.1	-5.9
35	S	6	15.0	0.40	11.8	20.97	-9.1	-5.8
36	M	15	15.4	0.97	16.1	20.94	-4.9	-1.1
37	T	24	15.4	1.56	19.1	20.92	-1.8	4.9
38	O	21	15.4	1.36	18.6	20.90	-2.3	2.4
39	T	18	15.4	1.17	12.9	20.87	-8.0	5.1
40	F	15	15.0	1.00	12.7	20.85	-8.1	2.3
41	L	9	13.7	0.66	14.9	20.83	-5.9	-5.9
42	S	6	12.0	0.50	11.8	20.80	-9.0	-5.8
43	M	12	15.7	0.76	12.9	20.78	-7.9	-0.9
44	T	15	15.3	0.98	12.0	20.76	-8.8	3.0
45	O	9	15.3	0.59	8.0	20.73	-12.8	1.0
46	T	44	16.9	2.61	31.5	20.71	10.8	12.5
47	F	12	18.9	0.64	10.2	20.69	-10.5	1.8
48	L	9	20.0	0.45	14.9	20.66	-5.8	-5.9
49	S	17	25.3	0.67	33.6	20.64	12.9	-16.6
50	M	26	21.4	1.21	27.8	20.62	7.2	-1.8
51	T	23	24.3	0.95	18.4	20.59	-2.2	4.6
52	O	46	24.6	1.87	40.8	20.57	20.2	5.2
53	T	17	23.7	0.72	12.2	20.55	-8.4	4.8
54	F	32	-	-	27.2	20.52	6.6	4.8
55	L	11	-	-	18.2	20.50	-2.3	-7.2
56	S	11	-	-	21.7	20.48	1.2	-10.7

Mellemresultater ved en opsplitning af en tidsrække i trend, sæsoneffekt og residual.

Ugedag	Gennemsnit af X_i/Z_i	Norm. gennemsnit = a_j
M	0.930	0.934
T	1.249	1.253
O	1.124	1.128
T	1.391	1.396
F	1.173	1.178
L	0.601	0.604
S	0.505	0.507
Sum	6.973	7.000

En måling taget på f.eks. en søndag korrigeres altså for ugevariationen ved, at den pågældende måling divideres med 0.507.

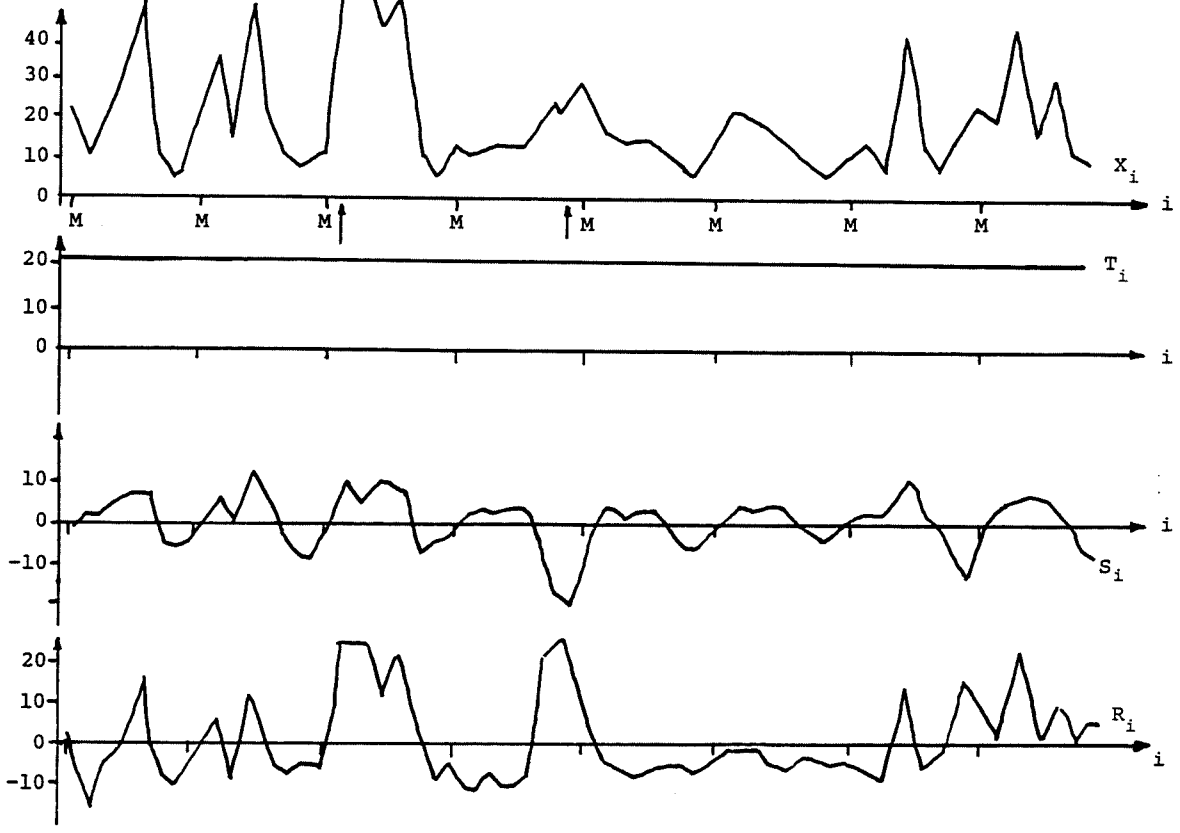
Trenden, der er bestemt ved almindelig lineær regression på værdierne af $T_i + R_i$, er fundet til

$$T_i = -0.0234\ i + 21.8 \ .$$

Ved at se på kurven R_i ser vi nu f.eks., at målingen taget den fjerde søndag (mærket med en †) er en lige så "extrem" måling som f. eks. observationen taget den tredie tirsdag (også mærket med en †), selv om denne sidste er ca. dobbelt så stor som den første.

Denne forskellige vurdering af tallene hænger som ovenfor nævnt sammen med, at vi er "vant til" relativt små målinger på søndage og til relativt store målinger midt på ugen. Derfor virker den kun middelstore måling den tredie søndag "dobbelt stærkt".

□



Observationen X_i er spaltet op i trenden T_i, i sæsoneffekten S_i og residual R_i .

9.4 Endimensional spektralanalyse

Vi skal i dette afsnit foretage en analyse af stationære tidsrækker i frekvensdomænet eller mere præcist foretage en analyse af Fourier-transformationerne af disse tidsrækkers autokovariansfunktioner. Der er mange lighedspunkter med Fourier-analysen af deterministiske signaler, men næsten selvsagt også afgørende forskelle. Som antydet forudsætter vi overalt i dette afsnit, at de involverede tidsrækker i det mindste er svagt stationære af anden orden.

Vi anfører først en række definitioner i

9.4.1 Spektret for en stokastisk proces

Det helt grundlæggende begreb i dette afsnit er spektret for en proces.

Definition 9.13 Ved spektret Γ_{XX} for en stationær stokastisk proces $X(t)$ forstås autokovariansfunktionen γ's Fourier-transformerede, i.e.

$$\Gamma_{XX}(f) = \int_{-\infty}^{\infty} \gamma_{XX}(u)\, e^{-i2\pi fu}\, du \ , \ -\infty < f < \infty \ ,$$

for en kontinuert proces og

$$\Gamma_{XX}(f) = \Delta \sum_{k=-\infty}^{\infty} \gamma_{XX}(k)\, e^{-i2\pi kf\Delta} \ , \ -\frac{1}{2\Delta} \leq f \leq \frac{1}{2\Delta} \ ,$$

for en proces, der observeres med tidsmellemrum Δ. I angelsaksisk litteratur benævnes spektret ofte "power spectrum".

Hvis man skal sammenligne spektre for forskellige processer, vil det oftest være hensigtsmæssigt at normere spektrene med variansen. Derved fremkommer den såkaldte spektraltæthed

$$\frac{\Gamma_{XX}(f)}{\sigma_X^2} \quad ,$$

og det ses, at denne simpelt hen er den Fourier-transformerede til autokorrelationsfunktionen. Ordet spektraltæthed bruges dog ofte også om $\Gamma_{XX}(f)$.

Ved hjælp af inversionsformlen finder vi i det kontinuerte tilfælde

$$\gamma_{XX}(u) = \int_{-\infty}^{\infty} \Gamma_{XX}(f)\ e^{i2\pi fu}\ df \ ,$$

og i det diskrete

$$\gamma_{XX}(k) = \int_{-1/(2\Delta)}^{1/(2\Delta)} \Gamma_{XX}(f)\ e^{i2\pi fk\Delta} df \ .$$

Sætter vi her u (eller k) lig 0, fås

$$\gamma_{XX}(0) = \sigma_X^2 = \int_{-\infty}^{\infty} \Gamma_{XX}(f)\ df \ ,$$

d.v.s. Γ_{XX} viser, hvorledes variansen af X(t)-processen er fordelt efter frekvenser. Specielt er den del af variansen af X(t)-processen, der skyldes frekvenser i området [f,f+df] approximativt lig med $\Gamma_{XX}(f)\ df$.

Ofte er man interesseret i, hvor stor en del af variansen, der skyldes frekvenser mindre end en vis værdi, og man ledes da til at definere det integrerede spektrum

$$I_{XX}(f_0) = \int_{-f_0}^{f_0} \Gamma_{XX}(f)\ df$$

Det fremgår trivielt, at

$$I_{XX}(0) = 0$$

og

$$I_{XX}(\infty) = \sigma_X^2$$

Beslægtet med det integrerede spektrum er den såkaldte spektralfordeling som defineret nedenfor.

Definition 9.14 Spektralfordelingen for en stationær proces med spektrum Γ_{XX} defineres ved

$$F_{XX}(f_0) = \int_{-\infty}^{f_0} \Gamma_{XX}(f)df$$

subsidiært

$$F_{XX}(f_0) = \int_{-\frac{1}{2\Delta}}^{f_0} \Gamma_{XX}(f)df$$

for henholdsvis en proces, der observeres til kontinuert tid, og en proces, der observeres med tidsmellemrum Δ.

Hvis man vil undgå at definere spektralfordelingen ud fra spektret, der muligvis kræver inddragelse af δ-funkiioner for at være defineret, kan man i stedet formulere inversionsformlen som et Stieltjes-integral, d.v.s.

$$\gamma_{XX}(u) = \int_{-\infty}^{\infty} e^{i2\pi fu} \, d \, F_{XX}(f)$$

respektive

$$\gamma_{XX}(k) = \int_{-\frac{1}{2\Delta}}^{\frac{1}{2\Delta}} e^{i2\pi fk\Delta} \, d\, F_{XX}(f)$$

i tilfældet med "kontinuert" henholdsvis "diskret" tid. Det kan vises, at der for en stationær proces altid eksisterer en ikke-aftagende funktion F_{XX}, som tilfredsstiller ovenstående. Ved især teoretiske undersøgelser kan det være en fordel at definere spektralfordelingen på denne måde.

Vi vil nu finde spektre for nogle af de processer, vi har betragtet.

Eksempel 9.26 (Spektrum for hvid støj). Hvis $A(t)$ er en kontinuert hvid støj, fås spektret

$$\Gamma_{AA}(f) = \int_{-\infty}^{\infty} \sigma^2 \, \delta(u) e^{-i2\pi fu} du = \sigma^2 \ ,$$

dvs. konstant lig variansen.

I det diskrete tilfælde fås

$$\Gamma_{AA}(f) = \Delta \sum_{k} \gamma(k) e^{-i2\pi fk\Delta} = \sigma^2\Delta \ ,$$

dvs. konstant lig variansen gange samplingafstanden. □

Vi betragter dernæst de p. 9.49 anførte MA(1)-processer.

Eksempel 9.27 Sætter vi

$$Z_t = \mu + A_t + \theta_1 A_{t-1} \ ,$$

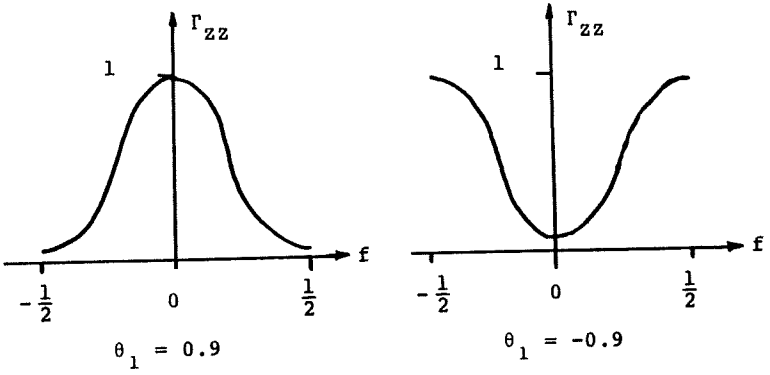
er som anført p. 9.49 med $\sigma_A^2 = 1$

$$\gamma_{ZZ}(k) = \begin{cases} 1+\theta_1^2 & k=0 \\ \theta_1 & k=1,-1 \\ 0 & \text{ellers} \end{cases}$$

Med $\Delta = 1$ bliver spektret

$$\Gamma_{ZZ}(f) = \theta_1 e^{+i2\pi f} + 1 + \theta_1^2 + \theta_1 e^{-i2\pi f}$$
$$= 1 + \theta_1^2 + 2\theta_1 \cos 2\pi f \quad , \quad |f| \leq \tfrac{1}{2} .$$

Denne funktion er for $\theta_1 = 0.9$ og -0.9 anført i nedenstående figur.



Disse spektre bør sammenlignes med graferne for de realiserede udfald af processerne, som er anført p. 9.50. For $\theta_1 = 0.9$ skyldes den overvejende del af variansen svingninger med lave frekvenser, og for $\theta_1 = -0.9$ er det de højere frekvenser, der er mest afgørende. Der ses at være en smuk overensstemmelse mellem det visuelle indtryk, man får af tidsrækkerne, og så de konklusioner, der drages på grundlag af spektrene. □

Vi betragter dernæst et eksempel vedrørende en autoregressiv proces.

Eksempel 9.28 Vi betragter de p. 9.54 anførte processer givet ved

$$Z_t = 2 + 0.9\, Z_{t-1} + A_t \ ,$$

og $$X_t = 2 - 0.9\, X_{t-1} + A_t \ .$$

Med $\phi_1 = 0.9$ henholdsvis -0.9 er autokovariansfunktionerne givet ved

$$\gamma(k) = \frac{1}{1-\phi_1^2}\, \phi_1^{|k|}, \quad k = 0, \pm 1, \pm 2, \cdots \ .$$

Af definitionen fås umiddelbart

$$\Gamma(f) = \frac{1}{1-\phi_1^2} \sum_{k=-\infty}^{\infty} \phi_1^{|k|}\, e^{-i2\pi fk} \quad , \quad -\tfrac{1}{2} \le f \le \tfrac{1}{2} \ .$$

Ved at spalte summen i led svarende til positive og negative k fås

$$\Gamma(f) = \frac{1}{1-\phi_1^2}\left[\frac{1}{1-\phi_1 e^{-i2\pi f}} + \frac{\phi_1 e^{i2\pi f}}{1-\phi_1 e^{i2\pi f}}\right]$$

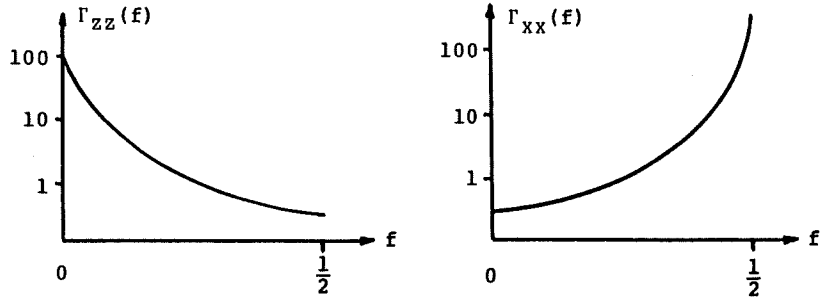
$$= \frac{1}{1+\phi_1^2-2\phi_1 \cos(2\pi f)} \ .$$

Vi har altså

$$\Gamma_{zz}(f) = \frac{1}{1.81-1.8\cos(2\pi f)} \ , \qquad -\tfrac{1}{2} \le f \le \tfrac{1}{2} \ ,$$

og $$\Gamma_{xx}(f) = \frac{1}{1.81+1.8\cos(2\pi f)} \ , \qquad -\tfrac{1}{2} \le f \le \tfrac{1}{2} \ .$$

Da disse funktioner har et variationsområde mellem 100 og 1/1.81, vil det være mest hensigtsmæssigt at afbilde dem i en logaritmisk skala. Dette er gjort i nedenstående figur.



Da funktionerne er symmetriske, har vi kun anført højre halvdel af grafen.

Det ses, at for X-rækkens vedkommende er den overvejende del af variation koncentreret i lave frekvenser svarende til meget langsomme svingninger. Y-rækkens varians er overvejende koncentreret i høje frekvenser svarende til meget hurtige svingninger. Dette er i pæn overensstemmelse med graferne for de realiserede udfald anført p. 9.54.

□

Inden vi betragter spektrene for de 3 processer anført i eksemplerne 9.26-28, må vi anføre nogle generelle resultater om spektre for lineære processer.

Sætning 9.30 Vi betragter en stokastisk proces $Y(t)$, der kan fås som output fra et lineært system med inputproces $X(t)$. Overføringsfunktionen for systemet kaldes $h(u)$. Da gælder

$$\Gamma_{YY}(f) = |H(f)|^2 \Gamma_{XX}(f) , \qquad -\infty \leq f \leq \infty ,$$

hvis processerne er kontinuert, respektive

$$\Gamma_{YY}(f) = \Delta^{-1} |H(f)|^2 \Gamma_{XX}(f) , \qquad -\frac{1}{2\Delta} \leq f \leq \frac{1}{2\Delta} ,$$

hvis processerne er diskrete og samplede med tidsintervaller Δ. Her angiver $H(f)$ den Fourier-transformerede til overføringsfunktionen. Udtrykt verbalt udsiger sætningen, at spektret for output-processen fås ved at multiplicere input-processens spektrum med kvadratet på modulus af overføringsfunktionens Fourier-integral.

Bevis Autokovariansfunktionen for $Y(t)$ er (jfr. beviset for sætning 9.26, p. 9.60).

$$\gamma_{XX}(u) = \int_0^\infty \int_0^\infty h(v)h(w)\gamma_{ZZ}(u+v-w)\, dv\, dw .$$

Da $Y(t)$ har spektret

$$\Gamma_{YY}(f) = \int_{-\infty}^{\infty} \gamma_{YY}(u)\, e^{-i2\pi fu}\, du ,$$

fås resultatet i sætningen ved indsætning af udtrykket for $\gamma_{YY}(u)$ og dernæst foretage nogle ombytninger af integrationsordenen.

Resultatet angående tidsrækker, der er registrerede til diskrete tidspunkter, vises ganske analogt.

□

Ved hjælp af sætningen findes let spektret for en lineær proces. Vi anfører resultatet i nedenstående

Corollar Vi betragter lineære processer $Z(t)$ henholdsvis Z_t, der fremkommer som output fra et lineært system med frekvensresponsfunktion ψ og input lig en hvid støj $A(t)$ henholdsvis A_t. Da er spektret for den lineære proces

$$\Gamma_{ZZ}(f) = \sigma_a^2 |\Psi(f)|^2 \quad , \quad -\infty < f < \infty$$

respektive

$$\Gamma_{ZZ}(f) = \Delta^{-1} \sigma_a^2 |\Psi(f)|^2 \quad , \quad -\frac{1}{2\Delta} \leq f < \frac{1}{2\Delta} ,$$

hvor $\sigma_a^2 = V(A(t))$, respektive $V(A_t)$, og Δ er samplingafstanden for den diskrete proces, og

$$\Psi(f) = \int_0^\infty \psi(u) e^{-i2\pi fu}\, du$$

respektive

$$\Psi(f) = \Delta \sum_{k=0}^{\infty} \psi_k \, e^{-i2\pi fk\Delta} \ .$$

Bevis Triviel følge af foregående sætning.

Vi skal dernæst se på spektrene for eksemplerne 9.16 (om sinusprocessen), 9.17 (den skjulte sinusproces) og 9.18 (den for styrrede sinusproces).

Eksempel 9.29 Sinusprocessen har autokovariansfunktionen

$$\gamma_k = \cos \frac{\pi}{3} k = \cos 2\pi \cdot \frac{1}{6} \cdot k \ .$$

Af exempel 9.11 ledes man til antagelsen

$$\Gamma(f) = \frac{1}{2} \{ \delta(f - \tfrac{1}{6}) + \delta(f + \tfrac{1}{6}) \} \ ,$$

men dette resultat er kun vist i det kontinuerte tilfælde. Vi må derfor verificere det ved hjælp af formlen p. 9.89:

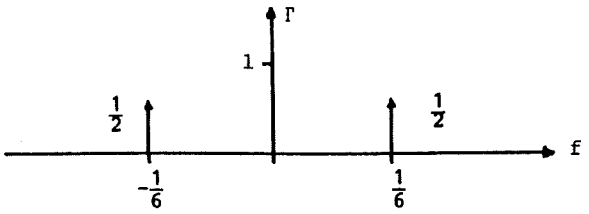
$$\int_{-\frac{1}{2}}^{\frac{1}{2}} \Gamma(f) e^{i2\pi fk} \, df = \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{1}{2} \{ \delta(f-\tfrac{1}{6}) + \delta(f+\tfrac{1}{6}) \} \, e^{i2\pi kf} \, df$$

$$= \frac{1}{2} \{ e^{i2\pi\frac{1}{6}k} + e^{-i2\pi\frac{1}{6}k} \}$$

$$= \cos \frac{\pi}{3} k$$

$$= \gamma_k \ ,$$

hvilket skulle vises. Grafen er skitseret nedenfor



□

Eksempel 9.30 Vi betragter den skjulte sinusproces

$$Y_t = Z_t + A_t \ .$$

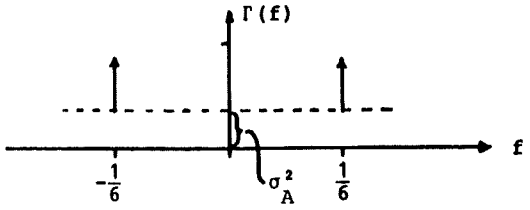
Da Y'er og Z'er er uafhængige, og da A er en hvid støj, fås

$$\begin{aligned}
\gamma_k &= \mathrm{Cov}(Y_t, Y_{t+k}) \\
&= \mathrm{Cov}(Z_t + A_t, Z_{t+k} + A_{t+k}) \\
&= \mathrm{Cov}(Z_t, Z_{t+k}) + \mathrm{Cov}(Z_t, A_{t+k}) + \mathrm{Cov}(A_t, Z_{t+k}) \\
&\quad + \mathrm{Cov}(A_t, A_{t+k}) \\
&= \begin{cases} \gamma_{ZZ}(k) & k \neq 0 \\ \gamma_{ZZ}(k) + V(A_t) \ , & k = 0 \end{cases} \\
&= \gamma_{ZZ}(k) + \gamma_{AA}(k) \ .
\end{aligned}$$

Følgelig er

$$\begin{aligned}
\Gamma(f) &= \Gamma_{ZZ}(f) + \Gamma_{AA}(f) \\
&= \tfrac{1}{2} \{ \delta(f - \tfrac{1}{6}) + \delta(f + \tfrac{1}{6}) \} + \sigma_A^2 \ .
\end{aligned}$$

Grafen er skitseret nedenfor



□

Eksempel 9.31 For den forstyrrede sinusproces er det vanskeligt at finde spektret direkte ud fra definitionen. I stedet benyttes, at X_t kan tolkes som output fra et lineært system med input A_t. Vi har altså

$$X(t) = \varphi\, X(t-1) - \varphi^2\, X(t-2) + A(t) \quad ,$$

d.v.s., systemet har frekvensresponsfunktionen

$$H(f) = \frac{1}{1 - \varphi e^{-i2\pi f} + \varphi^2 e^{-i4\pi f}}$$

ifølge sætning 9.10. Dette giver

$$\begin{aligned} |H(f)|^2 &= H(f)\ \overline{H(f)} \\ &= \frac{1}{1 - \varphi e^{-i2\pi f} + \varphi^2 e^{-i4\pi f}} \quad \frac{1}{1 - \varphi e^{i2\pi f} + \varphi^2 e^{i4\pi f}} \quad , \end{aligned}$$

d.v.s.

$$\begin{aligned} |H(f)|^{-2} &= 1 - \varphi e^{i2\pi f} + \varphi^2 e^{i4\pi f} - \varphi e^{-i2\pi f} + \varphi^2 \\ &\quad - \varphi^3 e^{i2\pi f} + \varphi^2 e^{-i4\pi f} - \varphi^3 e^{-i2\pi f} + \varphi^4 \\ &= 1 + \varphi^2 + \varphi^4 + \cos(2\pi f)\ \{-\varphi-\varphi-\varphi^3-\varphi^3\} \\ &\quad + \cos(4\pi f)\ \{\varphi^2+\varphi^2\} \\ &= 1 + \varphi^2 + \varphi^4 - 2\varphi(1+\varphi^2)\ \cos(2f) \\ &\quad + 2\varphi^2\ \cos(4\pi f) \quad . \end{aligned}$$

Følgelig er

$$\Gamma_{XX}(f) = \frac{1}{1 + \varphi^2 + \varphi^4 - 2\varphi(1+\varphi^2)\ \cos 2\pi f + 2\varphi^2\ \cos(4\pi f)} \quad .$$

Vi vil nu bestemme placeringen af eventuelle spidser. Vi finder

$$\begin{aligned} \frac{\partial}{\partial f}\, \Gamma_{XX}(f) &= \frac{-1}{T^2}\ (2\varphi(1+\varphi^2)\ 2\pi\ \sin(2\pi f) - 2\varphi^2 4\pi\ \sin(4\pi f)) \\ &= \frac{-1}{T^2}\ 4\varphi\pi\ \sin(2\pi f)\ \{1 + \varphi^2 - 4\varphi\ \cos(2\pi f)\} \quad , \end{aligned}$$

hvor T er tælleren i $\Gamma_{XX}(f)$. Vi har

$$\frac{\partial}{\partial f}\,\Gamma_{XX}(f) = 0 \quad \leftrightarrow \quad \cos 2\pi f = \frac{1+\varphi^2}{4\varphi} \ .$$

Skal dette være en spids, må fortegnsvariationen være

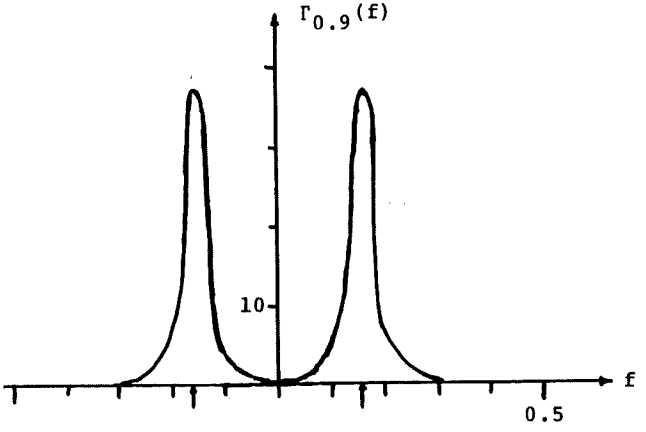
$$\xrightarrow[+\ +\ +\ +\ f_o\ -\ -\ -\ -]{\frac{\partial}{\partial f}\,\Gamma} \ ,$$

d.v.s. $1 + \varphi^2 < 4\varphi$ (cos ~ 1 for små f) .

For $\underline{\varphi = \frac{1}{2}}$ fås $\cos 2\pi f_o = \frac{1.25}{2} = 0.675$ eller $f_o \sim 0.132$.

For $\underline{\varphi = 0.9}$ fås $\cos 2\pi f_o = 0.5028$, d.v.s. $\underline{f_o \sim 0.166}$.

Da $\frac{1}{6} = 0.167$, ses, at spidsen for spektret i dette tilfælde ligger ret tæt på $\frac{1}{6}$ (og $-\frac{1}{6}$). Spektret er anført nedenfor.



□

Bemærkning Vi har i de tre sidste eksempler set, hvorledes forskellene i karakteren af rækkernes periodiciteter og dermed autokorrelationerne afspejler sig i spektrene.

9.4.2 Estimation af (power-)spektre

I dette afsnit betragter vi en proces X(t), der er observeret i tidsintervallet $-\frac{T}{2} \le t \le \frac{T}{2}$, respektive en proces, der er observeret til tidspunkter $-n\Delta,\cdots,0,\cdots,(n-1)\Delta$, og vi søger et estimat af spektret.

Da det teoretiske spektrum er Fourier-integralet af autokovariansfunktionen, vil det være åbenbart at starte med et estimat af denne, i.e.

$$c_{XX}(u) = \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} (x(t)-\bar{x})(x(t+u)-\bar{x})\,dt\,, \qquad -T \le u \le T\,,$$

respektive

$$c_{XX}(k) = \frac{1}{N}\sum_{t=-n}^{n-1-k} (x_t-\bar{x})(x_{t+k}-\bar{x})\,, \qquad k=0,1,\cdots,N-1\,,$$

hvor N = 2n. Man kan da naturligt definere

$$C_{XX}(f) = \int_{-T}^{T} c_{XX}(u)\, e^{-i2\pi fu}\,du\,, \qquad -\infty \le f \le \infty\,,$$

i det kontinuerte tilfælde, og

$$C_{XX}(f) = \Delta \sum_{k=-(N-1)}^{N-1} c_{XX}(k)\, e^{-i2\pi fk\Delta}\,, \quad -\frac{1}{2\Delta} \le f \le \frac{1}{2\Delta}\,,$$

i det diskrete tilfælde (jvf. de analoge formler p. 9.88).

Funktionen $C_{XX}(f)$ kaldes stikprøvespektret for den stokastiske proces.

Der gælder de sædvanlige inversionsformler, nemlig

$$c_{XX}(u) = \int_{\infty}^{\infty} C_{XX}(f)\, e^{i2\pi fu}\,df\,, \qquad -T \le u \le T\,,$$

respektive

$$c_{XX}(k) = \int_{-1/2\Delta}^{1/2\Delta} C_{XX}(f)\, e^{i2\pi fk\Delta} df \ , \qquad -N \le k \le N$$

For u eller k = 0 viser disse resultater i analogi med formlerne p. 9.89, hvorledes den empiriske varians er fordelt efter frekvenser.

Der gælder nu en yderst interessant sætning om sammenhængen mellem Fourier-integralet af x(t) respektive x_t og stikprøvespektret, nemlig

Sætning 9.31 Lad x(t) respektive x_t være realiserede udfald af en stokastisk proces og lad C_{XX} betegne stikprøvespektret. Da gælder

$$C_{XX}(f) = \frac{1}{T}\left|\int_{-\frac{T}{2}}^{\frac{T}{2}} x(t)\, e^{-i2\pi ft}\, dt\right|^2 \ , \qquad -\infty \le f \le \infty$$

respektive

$$C_{XX}(f) = \frac{\Delta}{N}\left|\sum_{k=-n}^{n-1} x_k\, e^{-i2\pi fk\Delta}\right|^2$$

$$= \frac{\Delta}{N}\left\{\left[\sum_{k=-n}^{n-1} x_k \cos 2\pi fk\Delta\right]^2 + \left[\sum_{k=-n}^{n-1} x_k \sin 2\pi fk\Delta\right]^2\right\} \ , \ -\frac{1}{2\Delta} < f < \frac{1}{2\Delta} \ .$$

Bevis Udmærket øvelse i transformering af integraler, respektive summer.

Bemærkning Sætningen viser, at stikprøvespektret, i.e. den empiriske autokovariansfunktions Fourier-transformation, er proportional med den kvadrerede amplitude af Fourier-transformationen af selve realisationen x. (Da realisationen x(t), respektive x_t, som regel er asymmetrisk, vil dets Fourier-transformation blive en komplex størrelse.

Vi skal nu se lidt på, i hvilken forstand stikprøvespektret er en rimelig estimator af de teoretiske spektre.

Vi betragter en normalt fordelt hvid støjproces A_t, der er observeret til tidspunkter $-n\Delta,\cdots,0,\cdots,(n-1)\Delta$, d.v.s., at A_t'erne er uafhængige og identisk $N(0,\sigma_A^2)$-fordelte. Vi sætter $2n = N$.

Vi har da spektret

$$C_{AA}(f) = \frac{\Delta}{N}\left[\left\{\sum_{t=-n}^{n-1} A_t \cos 2\pi ft\Delta\right\}^2 + \left\{\sum_{t=-n}^{n-1} A_t \sin 2\pi ft\Delta\right\}^2\right]$$

$$= \frac{\Delta}{N}\left[A^2(f) + B^2(f)\right], \qquad -\frac{1}{2\Delta} \le f \le \frac{1}{2\Delta}.$$

De fundamentale frekvenser benævnes

$$f_k = \frac{k}{N\Delta}, \quad k = 0,\pm 1,\cdots,\pm(n-1),-n.$$

Med disse benævnelser har vi følgende

Sætning 9.32 Idet vi i ovenstående situation sætter

$$Y(f_k) = \frac{2\, C_{AA}(f_k)}{\Delta\, \sigma_A^2}, \quad k = \pm 1,\cdots,\pm(n-1)$$

og

$$Y(f_k) = \frac{C_{AA}(f_k)}{\Delta\, \sigma_A^2}, \quad k = 0,-n$$

gælder

i) $Y(f_k) \in \chi^2(2)$, $k = \pm 1,\pm 2,\cdots,\pm(n-1)$

ii) $Y(f_k) \in \chi^2(1)$, $k = 0,-n$

iii) Alle $Y(f_k)$ er stokastisk uafhængige.

Bevis Det vises nogenlunde let, at $A(f_k)$ og $B(f_k)$ er normalt fordelte og ukorrelerede. Normalitet følger af, at der er tale om lineære transformationer af normalt fordelte variable. Ved at skrive kovariansmatricen op for de transformerede variable og benytte en ortogonalitetsegenskab ved de trigonometriske funktioner følger, at kovarianserne mellem forskellige $Y(f_k)$'er er nul.

Da

$$Y(f_k) = \frac{2}{n\sigma_A^2} [A^2(f_k) + B^2(f_k)] ,$$

følger resultatet let. □

Af sætningen fås trivielt følgende

Corollar Med ovenstående notation gælder

i) $E(C_{AA}(f_k)) = \sigma_A^2 \Delta = \Gamma_{AA}(f_k)$

og

ii) $V(C_{AA}(f_k)) = \sigma_A^4 \Delta^2 = \Gamma_{AA}^2(f_k)$.

Bemærkning 1 Det er overordentlig vigtigt at bemærke, at variansen på estimatet af spektret i de harmoniske frekvenser er uafhængigt af stikprøvestørrelsen. Dette bevirker altså, at der ikke direkte er vundet noget ved at forøge stikprøvestørrelsen. Nedenfor skal vi dog se, hvorledes man kan omgå dette problem (p. 9.106).

Bemærkning 2 En forudsætning for ovenstående resultater har været, at A_t-processen var normal. Hvis dette ikke er tilfældet, vil A(f) og B(f) dog alligevel være approximativt normale ifølge den centrale grænseværdisætning, og sætningernes resultater kan derfor opretholdes som approximative resultater.

I den næste sætning vil vi mere præcist angive forventningsværdien af stikprøvespektret.

Sætning 9.33 Lad C_{XX} betegne stikprøvespektret for en stokastisk proces X(t) (eller X_t) som defineret p. 9.100. Da gælder

$$E[C_{XX}(f)] = \int_{-T}^{T} \gamma_{XX}(u)\left(1-\frac{|u|}{T}\right) e^{-i2\pi fu}\, du ,$$

respektive

$$E[C_{XX}(f)] = \Delta \sum_{k=-(N-1)}^{N-1} \gamma_{XX}(k)\left(1-\frac{|k|}{N}\right) e^{-i2\pi fk\Delta} , \quad -\frac{1}{2\Delta} \le f \le \frac{1}{2\Delta} .$$

Bevis Forbigås.

Bemærkning Sættes

$$w_T(u) = \begin{cases} 1-\frac{|u|}{T} , & |u| \le T \\ 0 , & |u| > T \end{cases} ,$$

kan formlen for det kontinuerte tilfælde skrives

$$E[C_{XX}(f)] = \int_{-\infty}^{\infty} \gamma_{XX}(u) w_T(u)\, e^{-i2\pi fu}\, du ,$$

og af sætning 9.5 og eksempel 9.3 fås

$$E[C_{XX}(f)] = \int_{-\infty}^{\infty} T\left[\frac{\sin(\pi Tg)}{\pi Tg}\right]^2 \Gamma_{XX}(f-g)\, dg .$$

Dette resultat udtrykker, at stikprøvespektret har en forventet værdi, som svarer til at betragte det teoretiske spektrum gennem spektral-vinduet

$$W_T(f) = T\left[\frac{\sin(\pi Tf)}{\pi Tf}\right]^2 .$$

(eller mere stringent: forventningsværdien af $C_{XX}(f)$ er outputtet ved input $\Gamma_{XX}(f)$ fra et lineært system med overføringsfunktion W(f)).

Vi omtalte i bemærkning 1 til sætning 9.32 problemet med, at variansen på stikprøvespektret var uafhængig af stikprøvestørrelsen. Vi skal nu indføre nogle ændrede estimatorer, som bevirker en mindre varians.

Definition 9.15 Ved et lag-vindue $w(u)$ forstås en vilkårlig funktion, der tilfredsstiller

$$\text{i)} \quad w(0) = 1$$

$$\text{ii)} \quad w(u) = w(-u)$$

$$\text{iii)} \quad w(u) = 0, \ |u| > M, \quad M < T \ .$$

Ved det udglattede spektrum svarende til lag-vinduet w forstås i det kontinuerte henholdsvis det diskrete tilfælde

$$\overline{C}_{XX}(f) = \int_{-\infty}^{\infty} w(u) c_{XX}(u) \ e^{-i2\pi fu} \ du \ ,$$

respektive

$$\overline{C}_{XX}(f) = \Delta \sum_{k=-(M-1)}^{M-1} w(k) c_{XX}(k) \ e^{-i2\pi fk\Delta} \ .$$

Bemærkning Det udglattede spektrum svarer altså blot til at beregne spektret ud fra $w(u) \cdot c_{XX}(u)$ i stedet for ud fra kovariansfunktionen $c_{XX}(u)$. Benytter vi foldningssætningen, kan vi også skrive (i det kontinuerte tilfælde)

$$\overline{C}_{XX}(f) = \int_{-\infty}^{\infty} W(g) C_{XX}(f-g) \ dg \ ,$$

hvor $W(f)$ er Fourier-integralet til $w(u)$.

I nedenstående sætning har vi dernæst anført approximative formler for middelværdi og varians af det udglattede spektrum i det kontinuerte tilfælde.

Sætning 9.34 Lad notationen være som ovenfor. Da er

$$E[(\overline{C}_{XX}(f)] \simeq \int_{-\infty}^{\infty} W(g)\Gamma_{XX}(f-g)dg = \overline{\Gamma}_{XX}(f)$$

$$Cov[\overline{C}_{XX}(f_1),\ \overline{C}_{XX}(f_2)] \simeq \frac{1}{T}\int_{-\infty}^{\infty} \Gamma^2_{XX}(g)W(f_1-g)[W(f_2+g)+W(f_2-g)]\ dg\ .$$

Hvis $\Gamma^2_{XX}(g)$ er tilstrækkelig "glat", kan denne størrelse igen approximativt skrives

$$Cov[(\overline{C}_{XX}(f_1),\overline{C}_{XX}(f_2)] \simeq \frac{\Gamma^2_{XX}(f)}{T}\int_{-\infty}^{\infty} W(f_1-g)[W(f_2+g)+W(f_2-g)]\ dg\ .$$

Endvidere er

$$Var[\overline{C}_{XX}(f)] \simeq \frac{\Gamma^2_{XX}(f)}{T}\int_{-\infty}^{\infty} w^2(u)\ du = \Gamma^2_{XX}(f)\cdot\frac{I}{T}\ ,$$

hvor

$$I = \int_{-\infty}^{\infty} w^2(u)\ du\ .$$

Bevis Forbigås. En udmærket øvelse i approximation af integraler.

□

Bemærkning 1 Vi ser, at forventningsværdien af det udglattede spektrum approximativt er lig foldningen af det sande spektrum med spektralvinduet W. I afsnit 9.1.4 er effekten af en sådan "forstyrrelse" angivet. Det fremgår, at "afvigelsen" mellem den forstyrrede og den oprindelige funktion er mindst for store værdier af M (lag-vinduets bredde). Med andre ord er skævheden (bias)

$$B(f) = E(\overline{C}_{XX}(f)) - \Gamma_{XX}(f) \simeq \overline{\Gamma}_{XX}(f) - \Gamma_{XX}(f)$$

lille for store værdier af M. (M er defineret p. 9.105).

Bemærkning 2 Af udtrykket for kovariansen ses, at denne er proportional med graden af overlap af spektralvinduerne centreret i f_1 og f_2.

Bemærkning 3 Det fremgår, at variansen er stor for store værdier af M. For det rektangulære vindue fås e.g.

$$I = \int_{-M}^{M} 1^2 \, du = 2M ,$$

d.v.s. for dette vindue fås

$$\mathrm{Var}[(\overline{C}_{XX}(f)] \simeq \frac{\Gamma_{XX}^2(f)}{T} \cdot 2M .$$

Hvis vi ønsker en lille varians, må vi derfor gøre M lille. Dette er det stik modsatte af, hvad der var nødvendigt for at sikre en lille skævhed. Det endelige valg af M må derfor blive et kompromis mellem hensynene til skævheden og til variansen.

Det kan i øvrigt ofte være betydningsfuldt at gøre sig klart, hvorledes skævheder i estimationen af spektrene viser sig. I Parzen (1961) vises, at skævheden for såvel Hanning- som Parzen-vinduerne er af formen

$$B(f) = \frac{c}{M^2}\,\Gamma''_{XX}(f) + 0\left(\frac{1}{M^3}\right),$$

hvor c er en positiv konstant ($c \simeq 0.05 - 0.2$). Hvis Γ_{XX} har en "spids" i f_o, vil Γ''_{XX} være negativ i omegnen af f_o, og B(f) vil da være negativ, d.v.s. "spidser" tenderer til at blive underestimerede. Har Γ_{XX} derimod en "dal" i f_o, vil Γ''_{XX} være positiv i omegnen af f_o, og "dale" vil derfor fortrinsvis blive overestimerede.

Ved hjælp af sætning 9.34 kan vi som i bemærkning 3 finde approximative udtryk for variansen på spektre udglattede ved hjælp af standard-vinduerne.

Vi samler resultaterne i nedenstående tabel. Her er som tidligere anført T lig antal observationer og M den lag-værdi, uden for hvilken den udglattede autokorrelation er 0.

Navn	Spektralvindue	Variansforhold I/T
Rektangulære	$2M\,\frac{\sin(2\pi fM)}{2\pi fM}$	$2\,\frac{M}{T}$
Bartlett	$M\left[\frac{\sin(\pi fM)}{\pi fM}\right]^2$	$0.667\,\frac{M}{T}$
Hanning	$M\,\frac{\sin(2\pi fM)}{2\pi fM}\,\frac{1}{1-4f^2M^2}$	$0.75\,\frac{M}{T}$
Hamming	$M\,\frac{\sin(2\pi fM)}{2\pi fM}\,\frac{1.08-0.64f^2M^2}{1-4\,f^2M^2}$	$0.82\,\frac{M}{T}$
Parzen	$\frac{3}{4}\,M\left[\frac{\sin(\pi fM/2)}{\pi fM/2}\right]^4$	$0.539\,\frac{M}{T}$

Da som tidligere anført

$$V(\bar{C}_{XX}(f)) \simeq \Gamma^2_{XX}(f) \cdot \frac{I}{T},$$

kan vi ved hjælp af denne tabel finde ud af, hvor meget variansen på estimatet af spektret ændres ved at udglatte det.

Lad os f. eks. antage, at vi kan betragte kovariansfunktionen i intervallet [-0.1 T, 0.1 T], d.v.s. at M = 0.1 T. Da bliver ovenstående variansforhold

20%, 6.7%, 7.5%, 7.9% og 5.4%

for henholdsvis det rektangulære vindue, Bartlett-vinduet, Hanning-vinduet, Hamming-vinduet og Parzen-vinduet. Dette vil med andre ord sige, at variansen på det udglattede spektrum f. eks. er 6.7% af variansen på det "rå" stikprøvespektrum, hvis vi udglatter ved hjælp af et Bartlett-vindue.

Vi skal nu i et eksempel illustrere betydningen af at udglatte et spektrum. Eksemplet er taget fra Spliid (1973).

Eksempel 9.32 I vedstående figurer vises resultatet af en bølgemåling taget umiddelbart uden for Højer sluse. Endvidere er anført det rå spektrum, som viser en stor variabilitet. For at sikre et mere "roligt" estimat kan man foretage en udglatning, d.v.s. multiplicere autokorrelationsfunktionen $r(\tau)$ med en vægtfunktion

$$v(\tau) = \begin{cases} \left[\frac{1}{2} + \frac{1}{2}\cos(\tau \cdot \pi/128\tau_0)\right]^2 & , \quad \tau \leq 128\tau_0 \\ 0 & , \quad \tau > 128\tau_0 \end{cases}$$

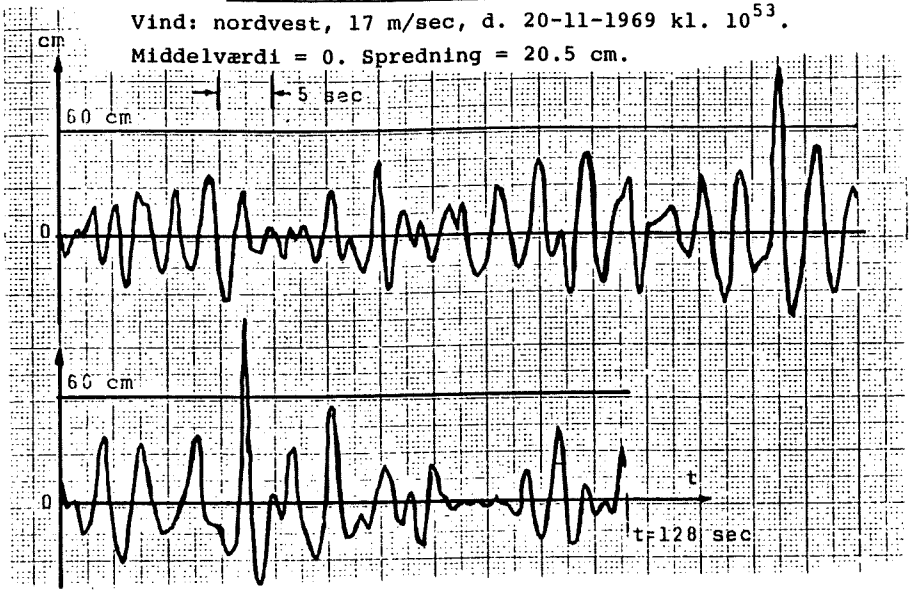
Denne er indtegnet på grafen over autokorrelationsfunktionen, og endelig er anført det udglattede spektrum. Man bemærker den store effekt af udglatningsproceduren.

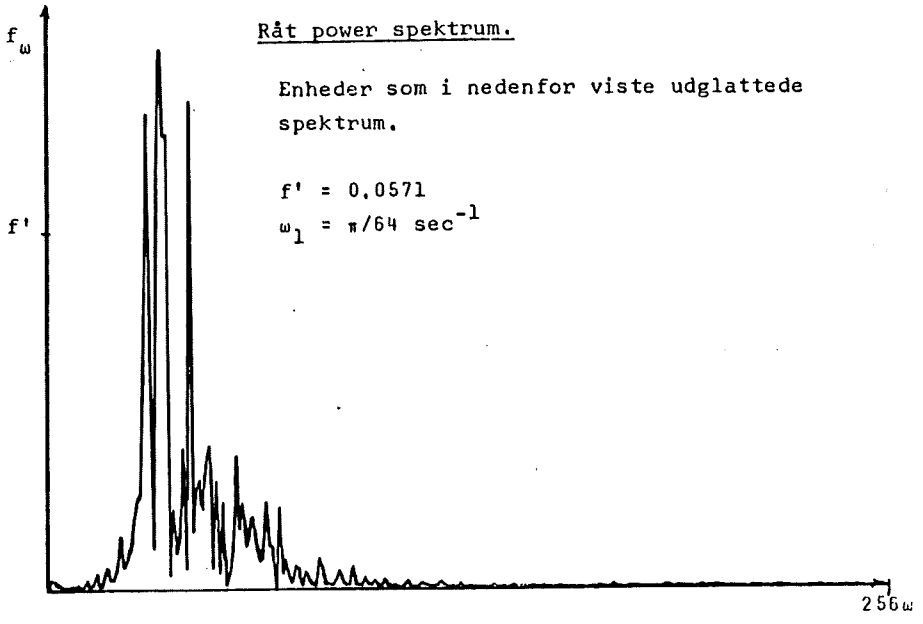
□

Bølgemåling fra Højer Sluse udv.

Vind: nordvest, 17 m/sec, d. 20-11-1969 kl. 10^{53}.

Middelværdi = 0. Spredning = 20.5 cm.



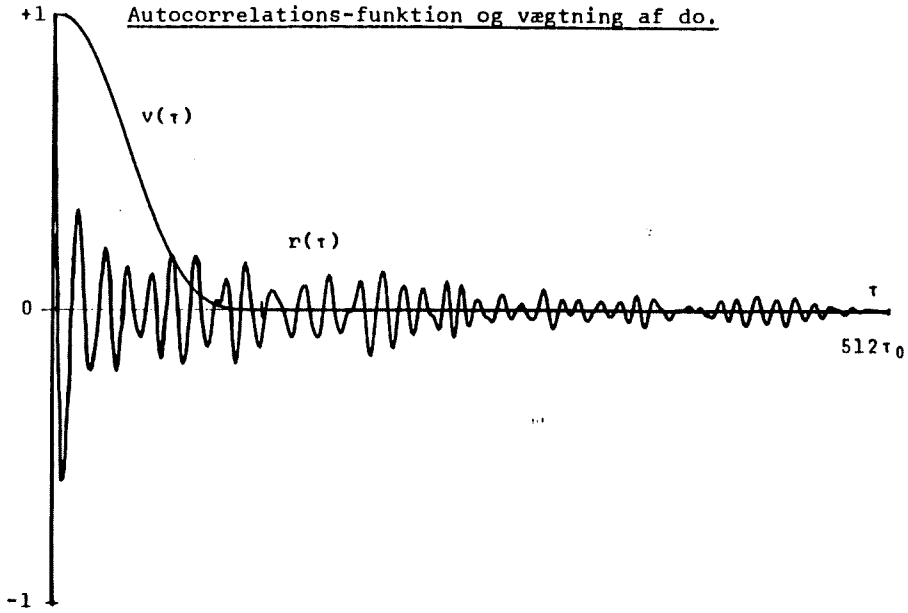


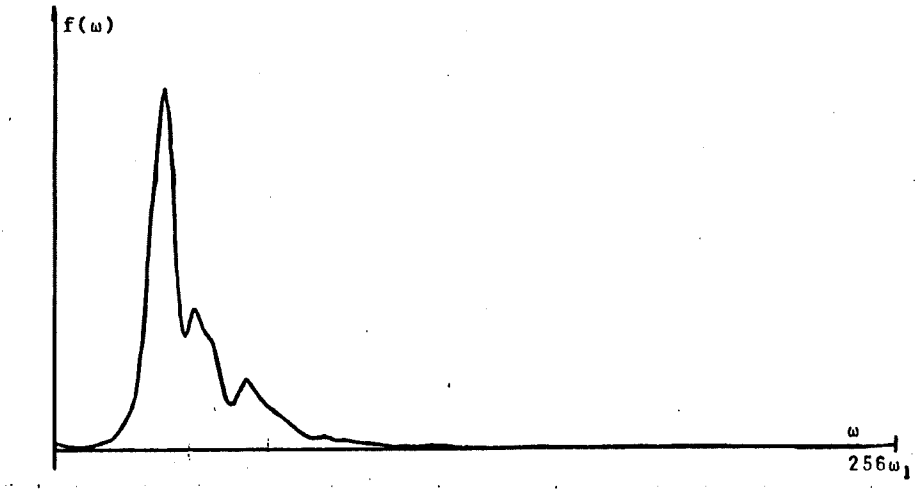
Råt power spektrum.

Enheder som i nedenfor viste udglattede spektrum.

$f' = 0{,}0571$

$\omega_1 = \pi/64 \text{ sec}^{-1}$


+1
Autocorrelations-funktion og vægtning af do.
v(τ)
r(τ)
0
τ
512τ0
-1


Udglattet power spektrum.
f(ω)
ω
256ω1

I sætning 9.32 fandt vi, at fordelingerne af stikprøvespektret var $\sigma^2\chi^2(2)$-fordelinger. Vi skal nu anføre de tilsvarende resultater for de udglattede spektre. Vi har

Sætning 9.35 Lad $\bar{C}_{XX}(f)$ betegne det udglattede spektrum svarende til lag-vinduet w (eller spektralvinduet W). Da gælder, at

$$\bar{C}_{XX}(f) \text{ approximativt } \in a\chi^2(\nu) \ ,$$

hvor

$$\nu = \frac{2[E(\bar{C}_{XX}(f))]}{V(\bar{C}_{XX}(f))} \ ,$$

$$a = \frac{E(\bar{C}_{XX}(f))}{\nu} \ .$$

Såfremt spektret er "glat" (med hensyn til vinduet jvf. sætning 9.34), kan ν og a approximeres ved

$$\nu \simeq 2T/\int_{-\infty}^{\infty} w^2(u)\ du = 2T/I$$

$$a \simeq \Gamma_{XX}(f)/\nu$$

Bevis Forbigås. Jvf. Jenkins & Watts (1968) p. 253.

□

Bemærkning Vi kan nu supplere skemaet p. 9.108 med en søjle indeholdende frihedsgradstallene svarende til de forskellige vinduer.

Vindue	Frihedsgradstal 2T/I
Rektangulære	T/M
Bartlett	3 T/M
Hanning	2.6667 T/M
Hamming	2.5164 T/M
Parzen	3.7086 T/M

Det ses umiddelbart, at dette giver mulighed for en væsentlig øgelse af frihedsgradsantallet fra de 2, som resultatet i sætning 9.32 anfører.

Ved hjælp af sætningen kan vi endvidere let angive konfidensintervaller for spektret. Vi anfører resultatet i følgende

Corollar Med den i sætning 9.35 anførte notation har vi

$$P\{ \chi^2(\nu)_{\alpha/2} \leq \frac{\nu\overline{C}_{XX}(f)}{\Gamma_{XX}(f)} \leq \chi^2(\nu)_{1-\alpha/2} \} \simeq 1-\alpha ,$$

eller

$$P\{ \frac{\nu\overline{C}_{XX}(f)}{\chi^2(\nu)_{1-\alpha/2}} \leq \Gamma_{XX}(f) \leq \frac{\nu\overline{C}_{XX}(f)}{\chi^2(\nu)_{\alpha/2}} \} \simeq 1-\alpha$$

Hvis vi afbilder spektrene i logaritmisk skala, kan vi anvende omskrivningen

$$P\{ \log \bar{C}_{xx}(f) + \log \frac{\nu}{\chi^2(\nu)_{1-\alpha/2}} \leq \log \Gamma_{xx}(f) \leq \log \bar{C}_{xx}(f) + \log \frac{\nu}{\chi^2(\nu)_{\alpha/2}} \} \simeq 1-\alpha$$

Bevis Triviel følge af foregående sætning.

□

Bemærkning Det fremgår, at det er en stor fordel at afbilde spektrene i logaritmisk målestok, thi da fås konfidensintervallerne ved overalt at addere henholdsvis subtrahere en konstant værdi fra det (udglattede) spektrum.

Vi skal nu anføre et lille eksempel til illustration af (nytten af) et udglattet power-spektrum.

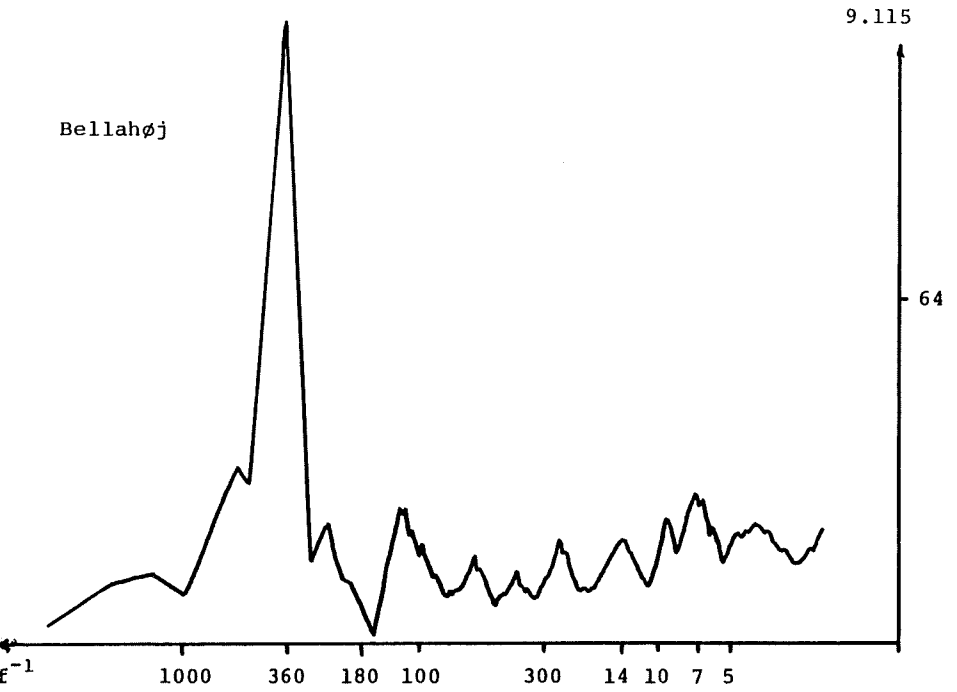
Eksempel 9.33 I Lyck & Gryning (1972) findes de p. 9.115 angivne power-spektre for en række døgnværdier af svævestøv målt på to forskellige lokaliteter i det storkøbenhavnske område.

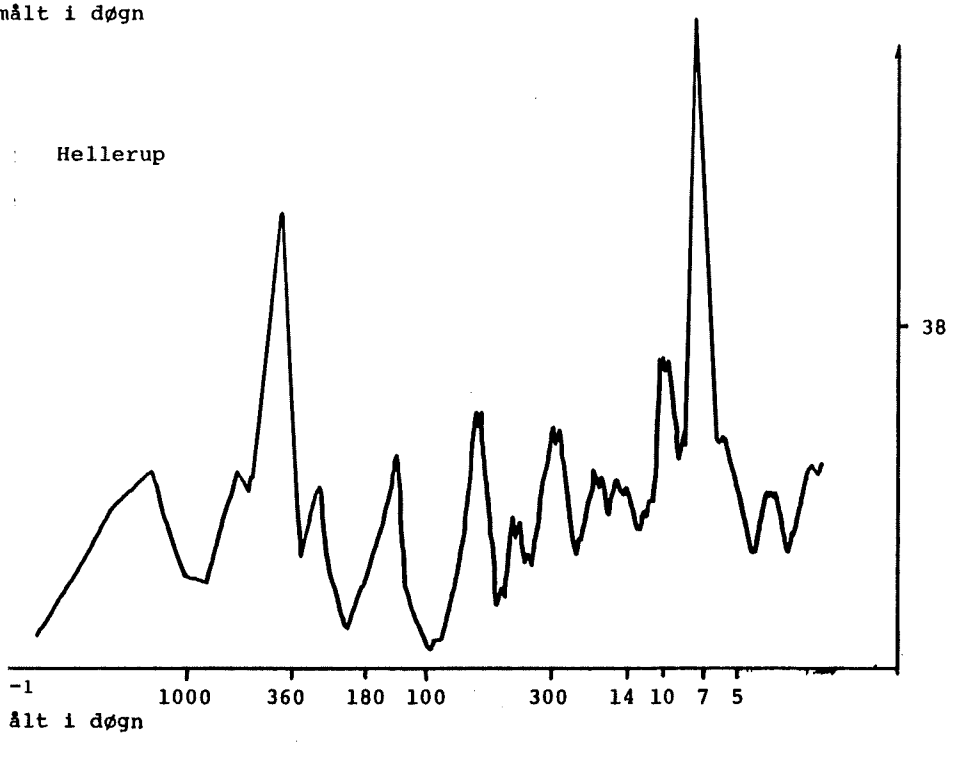
Mere specifikt drejer det sig om 4 års daglige bestemmelser af luftens gennemsnitlige indhold af svævestøv på to målestationer. Den ene station var placeret i Hellerup og den anden på Bellahøj.

Der er foretaget en spektralanalyse af dette materiale, i.e. der er bestemt et udglattet spektrum. Der er ikke angivet, hvilken udglatningsperiode der er anvendt.

Man bemærker, at begge tidsrækker har en meget stor del af variansen beliggende omkring frekvenser af størrelsesordenen 1/1 år. Der er med andre ord en væsentlig årsvariation i materialet - ikke noget overraskende resultat, problemstillingen taget i betragtning.

Mere bemærkelsesværdig er derimod den store forskel omkring frekvensen 1/1 uge. Her har spektret for Hellerup-målingerne en kraftig spids, hvorimod spektret for Bellahøj-målingerne ikke udviser nogen særlig markerede størrelsesforhold i dette område.


9.115
Bellahøj
64
f-1
1000
360
180
100
300
14
10
7
5
målt i døgn


Hellerup
38
-1
1000
360
180
100
300
14
10
7
5
ålt i døgn

Vi må derfor konkludere, at der er en tydelig ugevariation i støvindholdet ved Hellerup, hvorimod niveauet varierer mere tilfældigt og uafhængigt af ugedage på Bellahøj. Ved en nærmere analyse af luftkvaliteten i det storkøbenhavnske område vil det derfor være rimeligt at lede efter forhold, der kan betinge denne forskel i variationsmønsteret de to steder.

□

9.5 Filtrering

Ved en praktisk udførelse af en tidsrækkeanalyse viser det sig ofte, at rækken har den største del af variansen i ganske få frekvensområder. Dette vil ofte umuliggøre en mere detaljeret undersøgelse af andre frekvensområder. Det kan derfor være ønskeligt at "fjerne" disse frekvenser ved at "filtrere" data.

Der er sådan set intet principielt nyt i dette. Hvis man har et signal $x(t)$ og ønsker at fjerne frekvenser større end en grænse f_0, kan man blot multiplicere spektret $X(f)$ for $x(t)$ med (frekvens)vinduet

$$H_\ell(f) = \begin{cases} 1 & , \ |f| < f_0 \\ \frac{1}{2} & , \ |f| = f_0 \\ 0 & , \ |f| > f_0 \end{cases}$$

(ℓ for lav) og dernæst transformere det resulterende spektrum $X(f)H_\ell(f)$ "tilbage" ved hjælp af sætning 9.6. Man får da det filtrerede signal

$$y(t) = x_{\ell f_0}(t) = \int_{-\infty}^{\infty} h_\ell(u)x(t-u)\,du ,$$

hvor

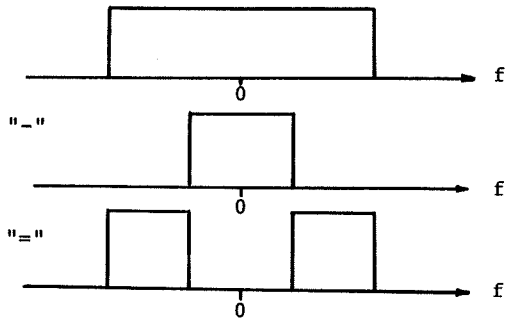
$$h_\ell(t) = 2f_0 \frac{\sin(2\pi f_0 t)}{2\pi f_0 t} .$$

Dette er det ideelle lav-pas filter, og det lader kun frekvenser mindre end f_0 passere. Det ideelle lav-pas filter svarer altså til "glidende linearkombinationer" (foldningsintegralet) med vægtene sin(x)/x.

Det er åbenbart, at man kan få et høj-pas filter (i.e. et filter, der lader frekvenser større end f_0 passere) ved simpelt hen at definere

$$x_{Hf_0}(t) = x(t) - x_{Lf_0}(t) \ .$$

Tilsvarende kan man definere bånd-pas filtre ved anvendelse af flere lav-pas filtre.



I praksis vil man dog ikke direkte kunne foretage en filtrering efter formlen

$$y(t) = \int_{-\infty}^{\infty} h_\ell(u)x(t-u)\ du \ ,$$

men må nøjes med approximationer som

$$y(t) \simeq y_1(t) = \int_{-T}^{T} h_\ell(u)x(t-u)\ du \ ,$$

hvor T angiver den tidshorisont, over hvilken "udjævningen" af

processen (signalet) foretages. Indfører vi det rektangulære (data)vindue

$$d_R(t) = \begin{cases} 1 & , \quad |t| < T \\ \frac{1}{2} & , \quad |t| = T \\ 0 & , \quad |t| > T \end{cases} ,$$

kan vi skrive

$$y_1(t) = \int_{-\infty}^{\infty} [h_\ell(u) d_R(u)] x(t-u)\, du$$
$$= \int_{-\infty}^{\infty} h_R(u) x(t-u)\, du ,$$

hvor altså

$$h_R(t) = h_\ell(t) d_R(t) .$$

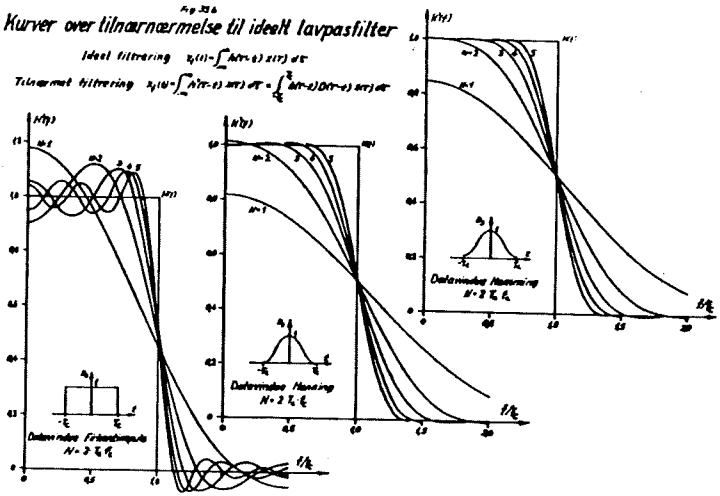
Det filter, som fremkommer på denne måde, er ikke særlig godt, så derfor vælger man som regel andre datavinduer ved multiplikationen med $h_\ell(t)$. Dette giver anledning til følgende impulsresponsfunktioner eller vægtfunktioner for filtrene

$$h_R(t) = h_\ell(t) d_R(t) \qquad \text{(rektangulære)}$$
$$h_B(t) = h_\ell(t) d_B(t) \qquad \text{(Bartlett)}$$
$$h_{Hn}(t) = h_\ell(t) d_{Hn}(t) \qquad \text{(Hanning)}$$
$$h_{Hm}(t) = h_\ell(t) d_{Hm}(t) \qquad \text{(Hamming)}$$
$$h_P(t) = h_\ell(t) d_P(t) \qquad \text{(Parzen)} ,$$

hvor $d_\nu(t)$, $\nu = R,B.Hn,Hm,P$ angiver de sædvanlige datavinduer (se f.eks. p. 9.5-9). Formlen for det med et datavindue $d_w(t)$ filtrerede signal bliver da

$$y_w(t) = \int_{-\infty}^{\infty} h_\ell(t) d_w(t) x(t-u) du = \int_{-\infty}^{\infty} h_w(t) x(t-u) du$$

I N.H. Hansen (1967) er bl.a. filtrene h_ℓ, h_{Hh} og h_{Hm} undersøgt, og vi anfører herfra følgende figur.



Det er helt åbenbart, at også andre funktioner kan komme på tale som filterfunktioner. De skal selvsagt blot have den egenskab, at de skal være "store" for de frekvenser, der skal "passere" filtret, og "små" for de frekvenser, der skal "tilbageholdes". Dette skal vi se eksempler på nedenfor, men først vil vi betragte problemet med diskrete signaler.

Vi <u>forudsætter</u> i det følgende, hvor intet andet er nævnt, at vi har et signal, der er <u>indsamlet med tidsafstanden Δ</u> .

Vi har da

<u>Sætning 9.36</u> Vi betragter et signal x(t), der forudsættes observeret til tiderne $i\Delta$, $-\infty < i < \infty$. Lad h være filterfunktion hørende til et kontinuert filter, og lad $h(t) = 0$ for $|t| > T$. Da er det digitale filter svarende til h givet ved

$$y(t) = y(i\Delta) = \Delta \sum_{j=-n}^{n} x((i-j)\Delta)h(j\Delta) \ .$$

Summationsgrænsen n er givet ved

$$T = n\Delta \ ,$$

(med åbenbare modifikationer, hvis T/Δ ikke er hel).

Bevis Forbigås, se f. eks. Dahlgaard (1973) p. 190.

Bemærkning 1 Sætter vi $y(i\Delta) = y_i$ og analogt med x og h, kan det filtrerede signal skrives

$$y_i = \Delta \sum_{j=-n}^{n} h_j \, x_{i-j}$$

Bemærkning 2 Ofte vil man være i stand til at filtrere et kontinuert signal, e.g. ad elektronisk vej, og dernæst sample det filtrerede signal til diskrete tidspunkter $i\Delta$, $i = \cdots,-1,0,1,\cdots$. Hvis Δ er så lille, at spektret X(f) for signalet x(t) er 0 for $|f| > 1/2\Delta$, da får man det samme resultat ved denne fremgangsmåde, som hvis man først havde samplet og dernæst filtreret, jvf. Dahlgaard (1973) p. 190.

Hidtil har vi ikke eksplicit udnyttet, at vi opfatter signalet eller tidsrækken x(t) som realisation af en stokastisk proces X(t). (Her må man ikke forveksle X(t) med spektret X(f) for signalet x(t)). Dette giver en drejning af resultaterne, og vi har

Sætning 9.37 (Digital filtrering for tidsrække) Lad X(t) være en stokastisk proces, der er observeret til tiderne $i\Delta$, $-\infty < i < \infty$. Vi sætter $X(i\Delta) = X_i$. For processen $Y_i = Y(i\Delta)$ defineret ved

$$Y_i = \sum_{\nu=-\infty}^{\infty} c_\nu \, X_{i-\nu}$$

$$= \cdots + c_{-n} \, X_{i+n} + \cdots + c_{-1} \, X_{i+1} + c_0 \, X_i + c_1 \, X_{i-1} + \cdots + c_n \, X_{i-n} + \cdots$$

gælder

$$\Gamma_{YY}(f) = |H(f)|^2 \, \Gamma_{XX}(f) \ ,$$

hvor

$$H(f) = \Delta \sum_{\mu=-\infty}^{\infty} c_\mu \, e^{-i2\pi f\mu\Delta} \ , \quad -\frac{1}{2\Delta} \leq f < \frac{1}{2\Delta} \ .$$

Bevis Forbigås. Rent teknisk.

Definition 9.16 Med notationen fra foregående sætning kaldes c_μ, $-\infty < \mu < \infty$, filtrets vægte eller vægtfunktion og $H(f)$ filtrets frekvensresponsfunktion.

Bemærkning Resultatet er fuldstændig analogt til sætning vedrørende spektret for en lineær proces. Den eneste forskel er, at vi her i udtrykket for Y_i også tillader fremtidige observationer at indgå.

Vi skal nu se på effekten af at foretage en udglatning af et filter (diskret) med vægte bestemt af et Hanning-vindue, jfr. de analoge betragtninger om kontinuerte filtre p. 9.118. Vi betragter et filter med vægte

$$\cdots, c_{-m-1}, c_{-m}, \cdots, c_0, \cdots, c_m, c_{m+1}, \cdots \ ,$$

og vi ønsker at afkorte dette, så det har længden $2m+1$. Dette kan trivielt gøres ved at sætte $c_\nu = 0$ for $|\nu| \geq m+1$; men som for de kontinuerte filtre vil det derved fremkomne filter ikke altid være tilfredsstillende. I stedet kan man betragte det filter, der fremkommer ved at anvende de filtervægte, der fremkommer af c_ν ved multiplikation med funktionen (Hanning-vinduet)

$$g(\nu) = \begin{cases} \frac{1}{2}(1 + \cos \frac{\pi\nu}{m+1}) & |\nu| \leq m+1 \\ 0 & \text{ellers} \ . \end{cases}$$

Bemærk, at $g(m+1) = 0$. I den angelsaksiske litteratur siger man, at d_ν fremkommer fra c_ν ved en Hanning-tapering. Vi har nu følgende sætning.

Sætning 9.38 (Hanning-tapering) Vi betragter et filter med vægtene c_ν, $-\infty < \nu < \infty$, og frekvensresponsfunktionen $H(f)$. Endelig betragtes frekvensresponsfunktionen H_m for filtret med vægte c_ν for $|\nu| \leq m$ og 0 for $|\nu| > m$, d.v.s.

$$H_m(f) = \Delta \sum_{\nu=-m}^{m} c_\nu e^{-i2\pi f\nu\Delta}, \quad -\frac{1}{2\Delta} \leq f < \frac{1}{2\Delta}.$$

Da har filtret med de ved en Hanning-tapering fremkomne vægte

$$d_\nu = \begin{cases} \frac{1}{2}\left(1 + \cos\frac{\pi\nu}{m+1}\right)c_\nu & |\nu| \leq m \\ 0 & \text{ellers} \end{cases}$$

frekvensresponsfunktionen

$$H_m^{Han}(f) = \frac{1}{4} H_m\left(f - \frac{1}{2(m+1)\Delta}\right) + \frac{1}{2} H_m(f) + \frac{1}{4} H_m\left(f + \frac{1}{2(m+1)\Delta}\right).$$

Hvis $c_\nu = 0$ for $|\nu| \geq m+1$, erstattes H_m i ovenstående med H.

Bevis Følger af definitionen ved anvendelse af formlen $\cos x = \frac{1}{2}\left(e^{ix} + e^{-ix}\right)$. Detaljerne forbigås.

Bemærkning Der gælder selvsagt et fuldstændig analogt resultat for en Hamming-tapering, d.v.s. en filtrering med vægtene

$$e_\nu = \begin{cases} \left(0.54 + 0.46 \cos\frac{\pi\nu}{m+1}\right)c_\nu & |\nu| \leq m \\ 0 & \text{ellers} \end{cases}$$

Da bliver frekvensresponsfunktionen

$$H_m^{Ham}(f) = 0.23\, H_m\left(f - \frac{1}{(2m+2)\Delta}\right) + 0.54\, H_m(f) + 0.23\, H_m\left(f + \frac{1}{(2m+2)\Delta}\right)$$

Vi vil nu se en række eksempler på digitale filtre. Vi vil endvidere antage, at $\Delta = 1$.

Eksempel 9.34 I en række standardprogrammer til tidsrækkeanalyse (e.g. BMD02T) kan man få udført en såkaldt prewhitening af ens data ved hjælp af et filter

$$Y_t = X_t - a\, X_{t-1} = (1-aB)X_t \ ,$$

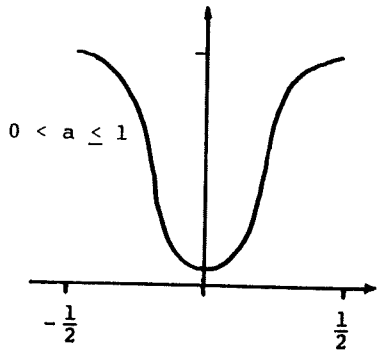
hvor a er en brugerspecificeret konstant med $|a| \leq 1$. Vi finder umiddelbart ved hjælp af sætning 9.10

$$H(f) = 1 - ae^{-i2\pi f}$$

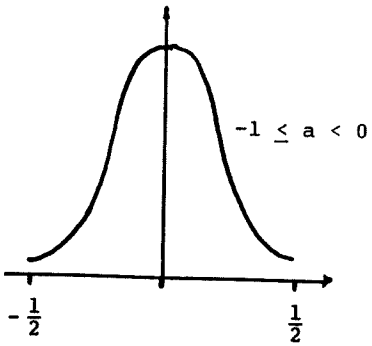
og dermed

$$|H(f)|^2 = 1 + a^2 - 2a\cos(2\pi f) \ .$$

I nedenstående figur er anført graferne for $|H(f)|^2$ for positive og negative værdier af a. Det ses, at filtret er et høj-pas filter for positive a og et lav-pas filter for negative a.



Høj-pas filter



Lav-pas filter

□

Eksempel 9.35 Vi betragter nu det filter, der fremkommer ved at tage glidende gennemsnit af observationerne, d.v.s. vi sætter

$$Y_t = \frac{1}{2m+1}\,(X_{t-m} + \cdots + X_t + \cdots + X_{t+m}) \ .$$

Vi får da

$$H(f) = \frac{1}{2m+1} \sum_{\mu=-m}^{m} \exp(-i2\pi\mu f)$$

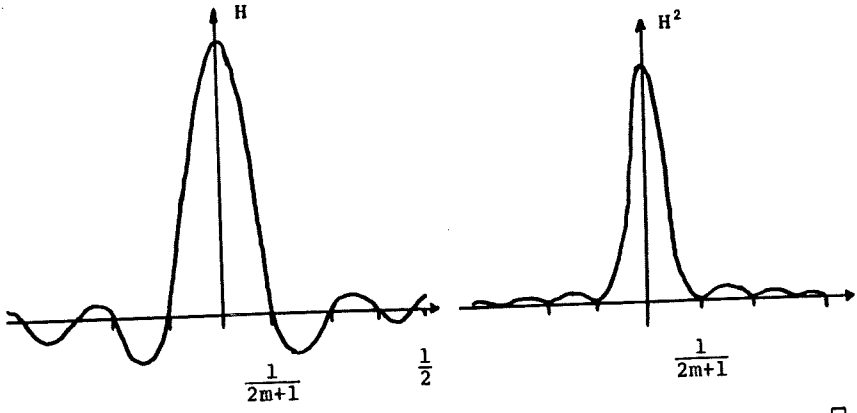
$$= \frac{1}{2m+1} \frac{\sin((2m+1)\pi f)}{\sin(\pi f)} \ , \qquad -\frac{1}{2} \le f < \frac{1}{2} \ ,$$

og dermed

$$|H(f)|^2 = \frac{1}{(2m+1)^2} \frac{\sin^2((2m+1)\pi f)}{\sin^2(\pi f)} \ , \qquad -\frac{1}{2} \le f < \frac{1}{2} \ .$$

Graferne for $H(f)$ og $|H(f)|^2$ er anført i nedenstående figur.

Det ses, at et glidende gennemsnit fungerer som et lav-pas filter, og at frekvenser større en 1/(længden af filtret) i det store og hele fjernes. Det fremgår dog også, at filtret ikke er særlig godt set i forhold til nogle af de tidligere (der er for mange relativt store lokale ekstrema mellem 1/(2m+1) og 1/2).



□

Eksempel 9.36 Da det ovenfor anførte filter ikke er helt tilfredsstillende som lav-pas filter, kan man søge at forbedre det ved en Hanning-tapering. Det giver filteret med vægtene

$$d_\nu = \begin{cases} \frac{1}{2m+1}\,\frac{1}{2}\left(1+\cos\frac{\pi\nu}{m+1}\right) & |\nu| \le m \\ 0 & \text{ellers} \end{cases} .$$

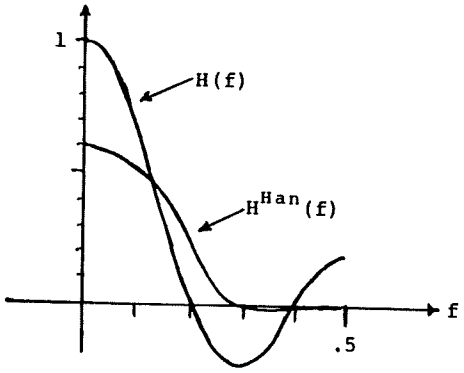
Sætter vi med notationen fra foregående eksempel

$$H(f) = \frac{1}{2m+1}\,\frac{\sin((2m+1)\pi f)}{\sin(\pi f)} \qquad -\tfrac{1}{2} \le f < \tfrac{1}{2} ,$$

bliver filtrets frekvensresponsfunktion

$$H^{Han}(f) = \tfrac{1}{4}\,H\left(f-\frac{1}{2m+2}\right) + \tfrac{1}{2}\,H(f) + \tfrac{1}{4}\,H\left(f+\frac{1}{2m+2}\right) .$$

Dennes graf er anført nedenfor i tilfældet m = 2.



Det ses, at der er opnået en væsentlig reduktion i forhold til H's lokale extrema.

□

Eksempel 9.37 Vi betragter nu et filter, der svarer til eksponentiel udjævning af første orden, i.e.

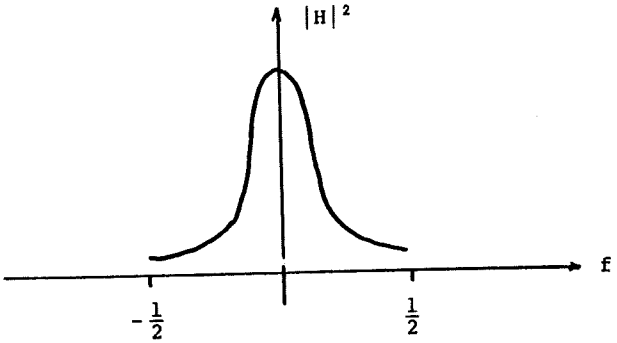
$$Y_t = \alpha\, Y_{t-1} + (1-\alpha) X_t \ ,$$

hvor $0 < \alpha \leq 1$. Vi får (jvf. p. 9.93)

$$H(f) = \frac{1-\alpha}{1-\alpha e^{-i2\pi f}} \ , \qquad -\frac{1}{2} \leq f < \frac{1}{2} \ ,$$

d.v.s.

$$|H(f)|^2 = \frac{(1-\alpha)^2}{1-2\alpha\ \cos(2\pi f)+\alpha^2} \ , \qquad -\frac{1}{2} \leq f < \frac{1}{2} \ .$$



Grafen for $|H(f)|^2$ er angivet ovenfor. Det ses, at filtret er et lav-pas filter.

Vi betragter nu i stedet eksponentiel udjævning svarende til en negativ udjævningskonstant. Vi får da

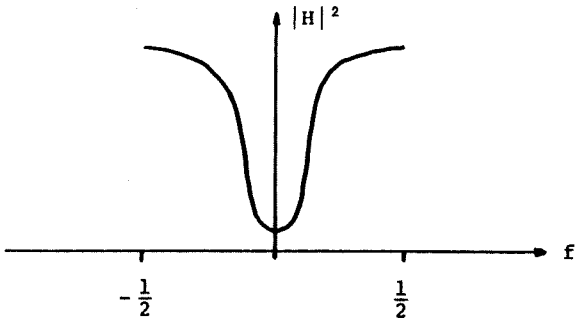
$$Y_t = \beta Y_{t-1} + (1+\beta) X_t \ ,$$

hvor $-1 \leq \beta < 0$. Da bliver

$$H(f) = \frac{1+\beta}{1-\beta e^{-i2\pi f}} \ , \qquad -\frac{1}{2} \leq f < \frac{1}{2}$$

og

$$|H(f)|^2 = \frac{(1+\beta)^2}{1-2\beta\ \cos(2\pi f)+\beta^2} \ , \qquad -\frac{1}{2} \leq f < \frac{1}{2}$$



Af ovenstående graf ses, at filtret er et høj-pas filter.

□

Ved analyse af økonomiske tidsrækker foretager man ofte en filtrering af data ved at danne differenser og summer af data. Dette kan give anledning til nogle tilsyneladende periodiciteter, som det fremgår af nedenstående eksempel.

Eksempel 9.38 Vi betragter to operatorer givet ved

$$(1-B)\ X_t = X_t - X_{t-1}$$

$$(1+B)\ X_t = X_t + X_{t-1}\ .$$

Det filter, der fås ved m sumdannelser og n differensdannelser, kan skrives

$$Y_t = (1+B)^m(1-B)^n\ X_t = \theta(B)\ X_t\ ,$$

hvor θ er et (n+m)'te grads polynomium.

Ved hjælp af sætning 9.10, p. 9.21, fås

$$H(f) = \theta(\exp(-i2\pi f)) = \left(1+e^{-i2\pi f}\right)^m \left(1-e^{-i2\pi f}\right)^n\ ,$$

og dermed

$$|H(f)|^2 = 2^{2n+2m}\ [\cos(\pi f)]^{2m}\ [\sin(\pi f)]^{2n}\ .$$

Man bringes let til at indse, at $|H(f)|^2$ har absolut maximum i

$$f = \frac{1}{2\pi} \text{Arccos}\left(\frac{1 - \frac{n}{m}}{1 + \frac{n}{m}}\right) .$$

Hvis man lader $m,n \to \infty$ på en sådan måde, at $n/m \to \theta$, kan det vises, at $|H(f)|^2$ konvergerer mod deltafunktioner placeret i f_0 og $-f_0$, hvor

$$f_0 = \frac{1}{2\pi} \text{Arccos}\left(\frac{1-\theta}{1+\theta}\right) .$$

Hvis X_t processen er hvid støj, er dens spektrum konstant. Spektret for den filtrerede proces vil derfor gå mod spektret for en cosinussvingning med frekvensen f_0 (se eksempel 9.11).

Vi har med andre ord frembragt en falsk periode f_0 ud fra den fuldstændig tilfældige proces X_t ved at foretage den nævnte filtrering. □

Bemærkning Resultatet i eksempel 9.38 er også kendt under navnet Slutski's sætning, og det vakte nogen opsigt ved sin fremkomst, og på denne måde lykkedes det at "forklare" nogle ellers "uforklarlige" periodiciteter i nogle filtrerede, økonomiske tidsrækker.

Vi betragter nu et konkret eksempel på filtrering af data.

Eksempel 9.39 I dette eksempel betragtes 320 målinger af "fejlen" (engelsk: "angle off") på et retursignal fra et radaranlæg, som følger et fly (eksemplet er baseret på Alavi & Jenkins (1965)). Disse fejl har forskellige årsager, bl.a. at flyet ikke optræder som en perfekt reflekterende kilde. Det resulterende signal modtaget af radaranlægget består derfor af en sammenblanding af signaler, som har vandret forskellige veje, og der er derfor tvivl om flyets sande position.

Af tekniske grunde er det ikke muligt at måle denne "target noise" uafhængigt af fejl forårsaget af flyets slingrebevægel-

ser. Som det fremgår af figuren p. 9.131, dominerer denne slingrebevægelse i det aktuelle tilfælde fuldstændigt den højfrekvente "target noise".

I det foreliggende tilfælde består tidsrækken altså af en blanding af to rækker, hvor det a priori vides, at den lavfrekvente del har en anden fysisk forklaring end den højfrekvente.

Vi vil nu søge at "skille" de to rækker ad ved en anvendelse af et passende valgt filter.

Som udgangspunkt vælger vi det i eksempel 9.36 anførte Hanning-filter.

Vi vil nu se, hvorledes filtret skal se ud, hvis vi ikke vil ændre på størrelsesforhold. Antallet af filtervægte benævnes $2m+1$. Da bliver vægtene

$$a_\nu = c\left(\frac{1}{2} + \frac{1}{2}\cos\frac{\pi\nu}{m+1}\right) \quad , \quad \nu = 0,\pm1,\cdots,\pm m \ ,$$

hvor c er en normeringskonstant, som skal bestemmes.

Da vi ønsker, at de filtrerede data skal være af samme størrelsesorden som de oprindelige, skal summen af a_ν'erne være 1. Vi får derfor

$$c \sum_{\nu=-m}^{m} \left(\frac{1}{2} + \frac{1}{2}\cos\frac{\pi\nu}{m+1}\right) = 1 \ ,$$

og ved hjælp af lidt regneri (eller en passende formelsamling) ses, at dette er ensbetydende med, at

$$c = \frac{1}{m+1} \ .$$

Vi har altså, at vægtene i lav-pas filtret er

$$a_\nu = \frac{1}{m+1}\left(\frac{1}{2} + \frac{1}{2}\cos\frac{\pi\nu}{m+1}\right) \quad , \quad = 0,\pm1,\cdots,\pm m \ .$$

Sættes det lav-pas filtrerede signal lig z_i, og definerer vi

$$\delta_{0\nu} = \begin{cases} 1 & \nu = 0 \\ 0 & \nu \neq 0 \end{cases}$$

(Kroneckers delta), kan vi skrive

$$x_i = \sum_{\nu=-m}^{m} \delta_{0\nu}\, x_{i-\nu}$$

$$z_i = \sum_{\nu=-m}^{m} a_\nu\, x_{i-\nu} \ ,$$

d.v.s. det høj-pas filtrerede signal $y_i = x_i - z_i$ kan skrives

$$y_i = \sum_{\nu=-m}^{m} \left(\delta_{0\nu} - a_\nu\right) x_{i-\nu}$$

$$= \sum_{\nu=-m}^{m} c_\nu\, x_{i-\nu} \ .$$

Vi har derfor følgende nyttige sammenhæng mellem koefficienterne a_ν i et lav-pas filter og koefficienterne c_ν i det tilsvarende høj-pas filter

$$c_\nu = \begin{cases} 1 - a_0 & \nu = 0 \\ - a_\nu & \nu \neq 0 \end{cases} \ .$$

I det konkrete tilfælde fås

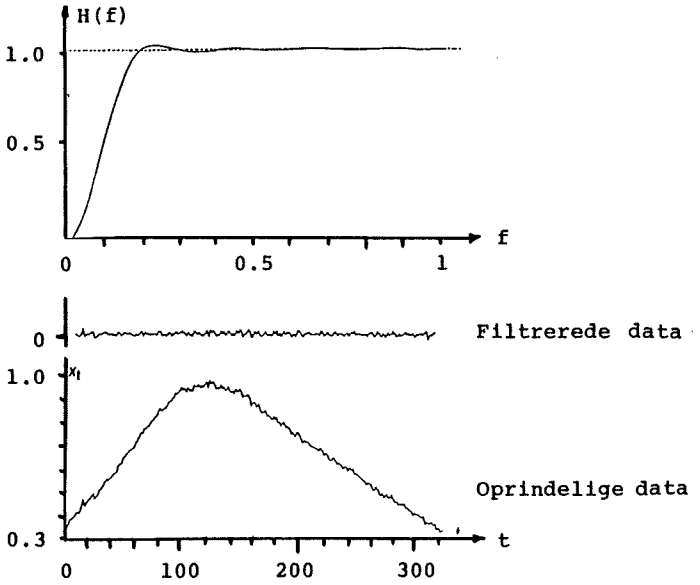
$$c_0 = 1 - \frac{1}{m+1}$$

$$c_\nu = - \frac{1}{m+1} \left(\frac{1}{2} + \frac{1}{2} \cos \frac{\pi\nu}{m+1}\right) \ .$$

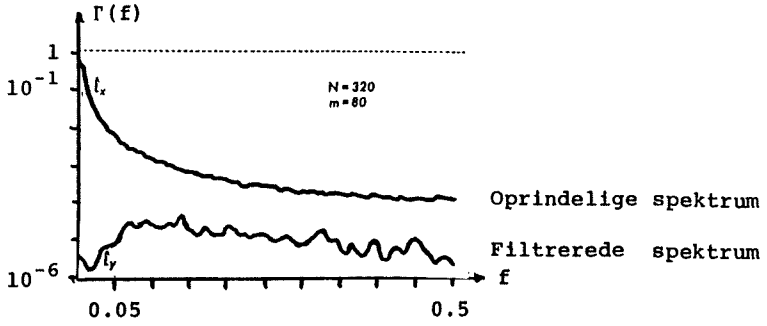
I det pågældende tilfælde er valgt m = 9, således at filtret har 19 led. I nedenstående figur er vist frekvensresponsfunktionen H(f) for dette filter.

Vi ser, at filtret udviser en smule for stor forstærkning omkring 0.2, men ellers synes det pænt. I den derpå følgende figur har vi vist både de filtrerede og de oprindelige data.

Der ses at være tale om en effektiv filtrering. Påvirkningen fra slingrebevægelsen i flyet er elimineret totalt.



Nedenfor er anført spektret for såvel den oprindelige som for den filtrerede proces. Spektret er beregnet på basis af et lagantal på 80, og udglatningen er foretaget ved hjælp af et Bartlett vindue.



Vi ser, hvorledes spektret for de oprindelige data er præget af den lavfrekvente del, og hvor det er fuldstændig umuligt at skelne detaljer i den højfrekvente del. Ved hjælp af filtreringen har vi fået focuseret på den i denne sammenhæng højfrekvente del af spektret, og den videre analyse kan nu foregå herudfra.

□

9.6 Krydsspektralanalyse

Ofte vil man ønske at analysere en vektorvariabels afhængighed af tiden. Analyserer man f. eks. strømmålinger (havstrømme), vil disse formentlig foreligge som en nordkomposant x(t) og en østkomposant y(t). En analyse af strømmens variation over tiden kræver da en samtidig analyse af

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix}, \quad t \in I,$$

hvor I er det relevante tidsrum. Et andet eksempel kan f. eks. være samtidige registreringer af luftens indhold af SO_2 på flere forskellige målelokaliteter. Her vil en analyse af forureningsmønsterets variation med tiden heller ikke kunne gøres alene ved marginale analyser af målingerne på de enkelte stationer. Der må også foretages en undersøgelse af samtidigheden eller tidsforskydningen mellem hændelser på de forskellige stationer.

Vi skal i dette afsnit alene give en indikation af de metoder, der kan tages i anvendelse. De fleste vil være centreret om analyser i frekvensdomænet. For en mere udførlig gennemgang også af tidsdomæne fremstillinger må henvises til f. eks. Jenkins & Watts (1968) eller den ret vanskeligt tilgængelige bog af Hannan (1970).

9.6.1 Krydskovarians og krydskorrelation

Vi vil starte med en ret summarisk gengivelse af de relevante definitioner og udskyde eksemplificeringen til det større eksempel i afsnit 9.6.5.

De fleste begreber i flerdimensional tidsrækkeanalyse bygger på begreber, der vedrører relationer mellem to tidsrækker. Vi betragter derfor i overvejende grad kun to tidsrækker

$$x_1(t) \quad \& \quad x_2(t) \ ,$$

der forudsættes at være realisationer af stationære stokastiske processer $X_1(t)$ og $X_2(t)$. Vi har da

$$E[X_i(t)] = \mu_i \ , \ i = 1,2$$
$$V[X_i(t)] = \sigma_i^2 \ , \ i = 1,2 \ ,$$

og autokovariansfunktionerne er givet ved

$$\gamma_{X_i X_i}(u) = \mathrm{Cov}[X_i(t), X_i(t+u)] = E[(X_i(t)-\mu_i)(X_i(t+u)-\mu_i] \ , \quad i = 1,2.$$

Disse angiver som tidligere anført kovarianserne mellem målinger af samme række taget med tidsmellemrum u.

Man har selvsagt også brug for kovarianserne mellem målinger på forskellige rækker taget med tidsmellemrum u. Derfor indføres krydskovariansfunktionerne

$$\gamma_{X_i X_j}(u) = \mathrm{Cov}[X_i(t), X_j(t+u)]$$
$$= E[(X_i(t)-\mu_i)(X_j(t+u)-\mu_j)]$$

Vi vil for simpelheds skyld anvende notationen

$\gamma_{ii}(u)$ i st. f. $\gamma_{X_i X_i}(u)$

og

$\gamma_{ij}(u)$ i st. f. $\gamma_{X_i X_j}(u)$.

Helt i analogi med indførelsen af autokorrelationsfunktionerne

$$\rho_{ii}(u) = \frac{\gamma_{ii}(u)}{\gamma_{ii}(0)} = \frac{\gamma_{ii}(u)}{\sigma_i^2} \quad , \quad i = 1,2 \ ,$$

indføres krydskorrelationsfunktionerne

$$\rho_{ij}(u) = \frac{\gamma_{ij}(u)}{\sqrt{\gamma_{ii}(0)\gamma_{jj}(0)}} = \frac{\gamma_{ij}(u)}{\sigma_i \sigma_j}$$

For autokorrelationerne og -kovarianserne gælder de sædvanlige relationer

$$\gamma_{ii}(u) = \gamma_{ii}(-u) \quad \& \quad \rho_{ii}(u) = \rho_{ii}(-u)$$

$$|\rho_{ii}(u)| \leq \rho_{ii}(0) = 1 \ .$$

d.v.s. de er symmetriske. Dette er ikke tilfældet for krydskovarianserne og -korrelationerne. Her gælder

$$\gamma_{12}(u) = \gamma_{21}(-u) \quad \text{og} \quad \rho_{12}(u) = \rho_{21}(-u) \ .$$

Disse relationer indses let. Endvidere gælder

$$|\rho_{ij}(u)| \leq 1 \ ,$$

som følger af, at

$$\forall \lambda,\mu: \ V[\lambda X_1(t) + \mu X_2(t+u)] = \lambda^2\sigma_1^2 + \mu^2\sigma_2^2 + 2\lambda\mu\sigma_1\sigma_2\rho_{ij}(u) \geq 0$$

(Ved et passende valg af (λ,μ) ses, at ovenstående relation netop giver det ønskede resultat).

Ofte vil det være nyttigt at skille krydskovariansfunktionen i en lige og en ulige del, d.v.s. vi indfører

$$\lambda_{12}(u) = \frac{1}{2}[\gamma_{12}(u) + \gamma_{12}(-u)]$$

$$\psi_{12}(u) = \frac{1}{2}[\gamma_{12}(u) - \gamma_{12}(-u)] \ ,$$

og har dermed

$$\lambda_{12}(-u) = \lambda_{12}(u) \quad \text{og} \quad \psi_{12}(-u) = -\psi_{12}(u)$$

$$\gamma_{12}(u) = \lambda_{12}(u) + \psi_{12}(u) \ .$$

Disse relationer kan anvendes såvel i det kontinuerte som i det diskrete tilfælde.

Vi vil nu betragte processer $X_1(t)$ og $X_2(t)$, der tilfredsstiller en relation som

$$X_2(t) = \int_0^\infty h(u) X_1(t-u)\, du + A(t) \ .$$

Her opfattes $X_1(t)$ altså som input til et lineært system og $X_2(t)$ som det af den uafhængige støj $A(t)$ forstyrrede output. Den diskrete analog til ovenstående er relationen

$$X_{2t} = \sum_{r=0}^{\infty} h_r X_{1t-r} + A_t \ .$$

Ifølge definitionen på krydskovariansfunktionen har vi - idet vi for simpelheds skyld forudsætter, at de involverede processer har middelværdi 0 -

$$\gamma_{12}(u) = E[X_1(t-u) \int_0^\infty h(v) X_1(t-v)\, dv + X_1(t-u)A(t)]$$

$$= \int_0^\infty h(v)\gamma_{11}(u-v)\, dv \ , \qquad -\infty < u < \infty \ ,$$

idet $\gamma_{X_1A}(u) = 0$ for alle u (grundet uafhængigheden). Tilsvarende fås

$$\gamma_{22}(u) = \int_0^\infty\int_0^\infty h(v)h(v')\gamma_{11}(u+v-v')\, dv\, dv' + \gamma_{AA}(u) \ ,$$

$$-\infty < u < \infty \ .$$

De diskrete analoger bliver

$$\gamma_{12}(k) = \sum_{r=0}^{\infty} h_r \gamma_{11}(k-r) \qquad k = 0,\pm1,\pm2,\cdots$$

$$\gamma_{22}(k) = \sum_{r=0}^{\infty}\sum_{s=0}^{\infty} h_r h_s \gamma_{11}(k+r-s) + \gamma_{AA}(k) \qquad k = 0,\pm1,\cdots .$$

9.6.2 Estimation af krydskovariansfunktion

Vi bestemmer i dette afsnit estimater, der er helt analoge til dem, der er anført i afsnittet om estimation af autokorrelationsfunktioner. I det kontinuerte tilfælde fås

$$c_{12}(u) = c_{X_1X_2}(u) = \begin{cases} \frac{1}{T}\int_{-T/2}^{T/2-u} (X_1(t)-\bar{X}_1)(X_2(t+u)-\bar{X}_2)\,dt , & 0 \le u \le T \\ \frac{1}{T}\int_{-T/2+u}^{T/2} (X_1(t)-\bar{X}_1)(X_2(t+u)-\bar{X}_2)\,dt , & -T \le u \le 0 \end{cases}$$

hvor

$$\bar{X}_i = \frac{1}{T}\int_{-T/2}^{T/2} X_i(t)\,dt , \quad i = 1,2 .$$

Hvis observationerne kommer fra en diskret tidsrække, d.v.s. hvis der foreligger målinger $X_{11},\cdots,X_{1N}$ og $X_{21},\cdots,X_{2N}$, bliver estimatorerne

$$c_{12}(k) = c_{X_1X_2}(k) = \frac{1}{N}\sum_{t=1}^{N-k} (X_{1t}-\bar{X}_1)(X_{2t+k}-\bar{X}_2) , \quad 0 \le k \le N-1 ,$$

$$c_{12}(-k) = c_{X_1X_2}(k) = \frac{1}{N}\sum_{t=1}^{N-k} (X_{1t+k}-\bar{X}_1)(X_{2t}-\bar{X}_2) , \quad 0 \le k \le N-1 ,$$

hvor

$$\bar{X}_i = \frac{1}{N}\sum_{t=1}^{N} X_{it} , \qquad i = 1,2 .$$

I Jenkins & Watts (1968) anføres formler for approximative værdier af momenter af disse estimater. Vi anfører (i det kontinuerte tilfælde)

Sætning 9.39 Lad situationen være som ovenfor. Da gælder

$$E[c_{12}(u)] = \left(1 - \frac{|u|}{T}\right)\gamma_{12}(u) + \frac{1}{T}\int_{-T}^{T}\left(1 - \frac{|t|}{T}\right)\gamma_{12}(t)\,dt$$

Bevis Forbigås.

Uden i øvrigt at komme ind på disse forhold skal vi præcisere, at estimaterne kan være stærkt korrelerede. Endvidere kan der "indføres" en "falsk" krydskorrelation, selv mellem ukorrelerede rækker på grund af autokorrelationer i de enkelte rækker.

Analogt til tidligere splittes også estimaterne for krydskovarianserne op i en lige og en ulige funktion, nemlig

$$\ell_{12}(u) = \frac{1}{2}\{\, c_{12}(u) + c_{12}(-u) \,\}$$

$$q_{12}(u) = \frac{1}{2}\{\, c_{12}(u) - c_{12}(-u) \,\},$$

formler, der er gyldige såvel i det diskrete som i det kontinuerte tilfælde.

9.6.3 Krydsspektret

Spektre indføres i det flerdimensionale tilfælde helt analogt til indførelsen i det endimensionale tilfælde. Det væsentlige begreb er krydsspektret, der defineres som Fourier-transformationen af krydskovariansfunktionen, d.v.s.

$$\Gamma_{12}(f) = \int_{-\infty}^{\infty} \gamma_{12}(u)\, e^{-i2\pi fu}\, du$$

i det kontinuerte tilfælde, og

$$\Gamma_{12}(f) = \Delta \sum_{k=-\infty}^{\infty} \gamma_{12}(k)\, e^{-i2\pi fk\Delta}, \qquad \frac{1}{2\Delta} \leq f < \frac{1}{2\Delta} .$$

Der optræder imidlertid en komplikation i forhold til analysen af autokovariansens Fourier-transformation. Autokovariansfunktionen er nemlig symmetrisk, hvorfor spektret bliver reelt. Dette er ikke tilfældet for krydskovariansen. Man kan derfor vælge forskellige former for angivelse af den komplekse funktion $\Gamma_{12}(f)$, nemlig dels en angivelse af modulus og argument (amplitude og fase) og dels en angivelse af realdel og imaginærdel. Dette giver

$$\begin{aligned} \Gamma_{12}(f) &= \alpha_{12}(f)\, e^{i\varphi_{12}(f)} \\ &= \Lambda_{12}(f) - i\Psi_{12}(f) , \end{aligned}$$

hvor

$$\alpha_{12}(f) = |\Gamma_{12}(f)| = \sqrt{\Lambda_{12}^2(f) + \Psi_{12}^2(f)}$$

kaldes kryds-amplitudespektret,

$$\varphi_{12}(f) = \operatorname{arctg}\left[-\frac{\Psi_{12}(f)}{\Lambda_{12}(f)}\right]$$

kaldes fasespektret, og

$$\Lambda_{12}(f) = \int_{-\infty}^{\infty} \lambda_{12}(u)\, e^{-i2\pi fu}\, du$$

og

$$\Psi_{12}(f) = \int_{-\infty}^{\infty} \psi_{12}(u)\, e^{-i2\pi fu}\, du$$

benævnes henholdsvis co-spektrum og kvadraturspektrum. De sidste formler er kun gyldige i det kontinuerte tilfælde. De diskrete

analoger bliver

$$\Lambda_{12}(f) = \Delta \sum_{k=-\infty}^{\infty} \lambda_{12}(k)\, e^{-i2\pi fk\Delta}, \quad -\frac{1}{2\Delta} \le f < \frac{1}{2\Delta},$$

og

$$\Psi_{12}(f) = \Delta \sum_{k=-\infty}^{\infty} \psi_{12}(k)\, e^{-i2\pi fk\Delta}, \quad -\frac{1}{2\Delta} \le f < \frac{1}{2\Delta}.$$

Det ses let, at $\alpha_{12}(f)$ er en lige funktion af f, og at $\varphi_{12}(f)$, $\Lambda_{12}(f)$ og $\Psi_{12}(f)$ er ulige funktioner af f.

Endelig indføres det kvadrerede koherensspektrum ved

$$\kappa_{12}^2(f) = \frac{\alpha_{12}^2(f)}{\Gamma_{11}(f)\Gamma_{22}(f)}.$$

Indførelsen af denne størrelse begrundes senere.

Vi vil nu se nærmere på det p. 9.135 indførte system givet ved

$$X_2(t) = \int_0^{\infty} h(u) X_1(t-u)\, du + A(t).$$

Da krydskovariansfunktionen er lig foldningen af h og γ_{11}, får vi af sætning 9.6

$$\Gamma_{12}(f) = H(f)\, \Gamma_{11}(f),$$

hvor H(f) er lig frekvensresponsfunktionen for det lineære system.

Sættes

$$H(f) = G(f)\, e^{i\varphi(f)},$$

hvor G altså er forstærkningen og φ fasefunktionen for det lineære system, fås derfor

$$H(f) = \frac{\Gamma_{12}(f)}{\Gamma_{11}(f)}$$

$$G(f) = \frac{\alpha_{12}(f)}{\Gamma_{11}(f)}$$

$$\varphi(f) = \text{arctg} - \frac{\psi_{12}(f)}{\Lambda_{12}(f)} \quad .$$

Disse formler angiver således, hvorledes det lineære systems karakteristika kan bestemmes ved hjælp af krydsspektret for input-output og spektret for input.

Betragter vi dernæst formlen for $\gamma_{22}(u)$, fås

$$\Gamma_{22}(f) = \int_{-\infty}^{\infty}\int_{0}^{\infty}\int_{0}^{\infty} h(v)h(v')\gamma_{11}(u+v'-v')\, e^{-i2\pi fu}\, dvdv'du + \Gamma_{zz}(f)$$

$$= H(f)H(-f)\Gamma_{11}(f) + \Gamma_{zz}(f)$$

$$= G^2(f)\Gamma_{11}(f) + \Gamma_{zz}(f) \quad .$$

Indsættes heri det ovenfor bestemte udtryk for G(f), fås

$$\Gamma_{zz}(f) = \Gamma_{22}(f) - \frac{\alpha_{12}^2(f)}{\Gamma_{11}(f)} = \Gamma_{22}(f)\ \{\ 1 - \kappa_{12}^2(f)\ \} \quad .$$

Af denne formel ses, at hvis støjspektret $\Gamma_{zz}(f)$ er lig outputspektret, er koherensen lig 0. Hvis $\Gamma_{zz}(f) = 0$, er den kvadrerede koherens lig 1.

Vi ser med andre ord, at koherensen svarer til en korrelationskoefficient defineret for hver frekvens f.

Til sidst i dette afsnit skal vi anføre en relation, der er analog til den p. 9.89 for spektret angivne egenskab. Vi har

$$\gamma_{12}(u) = \int_{-\infty}^{\infty} \Gamma_{12}(f)\, e^{i2\pi fu}\, df$$

$$= \int_{-\infty}^{\infty} \Lambda_{12}(f)\, \cos(2\pi fu)\, df - i \int_{-\infty}^{\infty} \Psi_{12}(f)\, \sin(2\pi fu)\, df ,$$

og dermed for $u = 0$

$$\boxed{\gamma_{12}(0) = \int_{-\infty}^{\infty} \Lambda_{12}(f)\, df ,}$$

det vil med andre ord sige, at co-spektret giver dekomponeringen af kovariansen svarende til lag 0 efter de forskellige frekvenser.

9.6.4 Estimation af krydsspektret

Et åbenbart forsøg på at estimere de forskellige komponenter i krydsspektret ville være at indsætte estimater for krydskovariansfunktionen og autokovariansfunktionerne i de teoretiske relationer mellem kovariansfunktioner og spektre. Det viser sig imidlertid, ganske som det var tilfældet ved estimationen af spektrene, at disse estimaters varianser ikke går mod 0 for stigende stikprøvestørrelser. Man må derfor skride til samme løsning her, som blev gjort tidligere, nemlig at tillade en vis "bias" for at få mindre varians. Dette gøres også her ved en udglatningsprocedure.

Vi betegner med $w(u)$ et lagvindue med egenskaber som anført i definitionen p. 9.105. Vi har da det udglattede skøn

$$\boxed{\bar{C}_{12}(f) = \int_{-M}^{M} w(u) c_{12}(u)\, e^{-i2\pi fu}\, du}$$

i det kontinuerte tilfælde og

$$\boxed{\bar{C}_{12}(f) = \Delta \sum_{k=-M}^{M} w(k) c_{12}(k)\, e^{-i2\pi fk\Delta}}$$

i det diskrete tilfælde (med åbenbare modifikationer i summationsgrænserne, hvis M ikke er heltallig). De øvrige spektre kan nu findes ved hjælp af definitionerne ud fra $\overline{C}_{12}$ (eller beregnes analogt direkte ud fra krydskorrelationsfunktionen).

Vi skal ikke her komme ind på særligt mange resultater vedrørende egenskaberne ved de udglattede estimater, men blot anføre

Sætning 9.40 Vi betragter det symmetriske lagvindue w(u) med $w(u) = 0$, $u > M$, og med $w(0) = 1$. Sættes

$$I = \int_{-M}^{M} w^2(u)\, du$$

gælder under passende glathedsbetingelser for de teoretiske spektre, at variansen på det udglattede krydsamplitudespektrum er

$$V[\overline{A}_{12}(f)] \simeq \frac{I}{2T}\, \alpha_{12}^2(f)\left[1 + \frac{1}{\kappa_{12}^2(f)}\right] ,$$

og variansen på det udglattede koherens- og det kvadrerede koherensspektrum er

$$V[|\overline{K}_{12}(f)|] \simeq \frac{I}{2T}\, [1 - \kappa_{12}^2(f)]^2 ,$$

henholdsvis

$$V[\overline{K}_{12}^2(f)] \simeq \frac{I}{2T}\, 4\kappa_{12}^2(1 - \kappa_{12}^2)^2 .$$

Endelig er variansen på det udglattede fasespektrum

$$V[\overline{F}_{12}(f)] \simeq \frac{I}{2T}\left[\frac{1}{\kappa_{12}^2(f)} - 1\right] .$$

Endelig gælder, at fase og krydsamplitude er approximativt ukorrelerede, og at fase og kvadreret koherens er approximativt ukorrelerede.

Bevis Forbigås. Se f. eks. Jenkins & Watts (1968).

Bemærkning 1 Vi ser, at varianserne på krydsamplitude- og fasespektret går mod ∞ for koherensen gående mod 0, d.v.s. man må være forberedt på, at ens estimater kan være stærkt påvirkede af en eventuelt forsvindende koherens.

Bemærkning 2 Ved hjælp af ovenstående sætning kan man konstruere kurveblade til beregning af konfidensintervaller for koherens- og fasespektrum. Disse kan findes i litteraturen.

Vi skal dernæst betragte den bias, der er forbundet med kohærensestimation.

Sætning 9.41 For Parzen- og for Hanning-vinduet gælder, at bias på det kvadrerede (udglattede) koherensspektrum approximativt er givet ved

$$B(f) \simeq \frac{a}{T}\frac{M}{} + \frac{b}{M^2}\left\{ \frac{\alpha_{12}(f)\alpha''_{12}(f) - \alpha^2_{12}(f)[\varphi'_{12}(f)]^2}{\Gamma_{11}(f)\Gamma_{22}(f)} \right\}.$$

For Parzen-vinduet gælder (a,b) = (0.54,0.304), og for Hanning-vinduet er (a,b) = (0.75,0.126).

Bevis Forbigås. Se f. eks. Jenkins & Watts (1968).

Bemærkning Vi ser, at bias er proportional med φ'_{12}. Hvis der derfor er store forsinkelser mellem de to processer (hvilket jo medfører, at $\varphi'_{12}(f)$ er stor), vil man få en stor bias i kohærensestimatet. Denne bias vil kunne elimineres ved en såkaldt "alignment", d.v.s. en forskydning i tiden af den ene proces, så krydskovariansfunktionen antager sin maximalværdi i 0.

9.6.5 Eksempel på krydsspektralanalyse

I dette eksempel vil vi måske ikke så meget koncentrere os om en direkte løsning af det problem, der behandles, men snarere søge at illustrere de begreber, der er indført i det tidligere.

Vi skal søge at illustrere sammenhængen mellem tre tidsrækker, nemlig

- $x(t)$: emission af SO_2 (i kg/½h) fra en punktkilde, d.v.s. den mængde svovldioxid, der udsendes fra kilden,
- $y(t)$: immission af SO_2 (i $\mu g/m^3$) på målelokalitet 1, 4 km fra kilden, d.v.s. den målte koncentration 4 km fra kilden,
- $z(t)$: immission af SO_2 (i $\mu g/m^3$) på målelokalitet 2, 8 km fra kilden, d.v.s. den målte koncentration 8 km fra kilden.

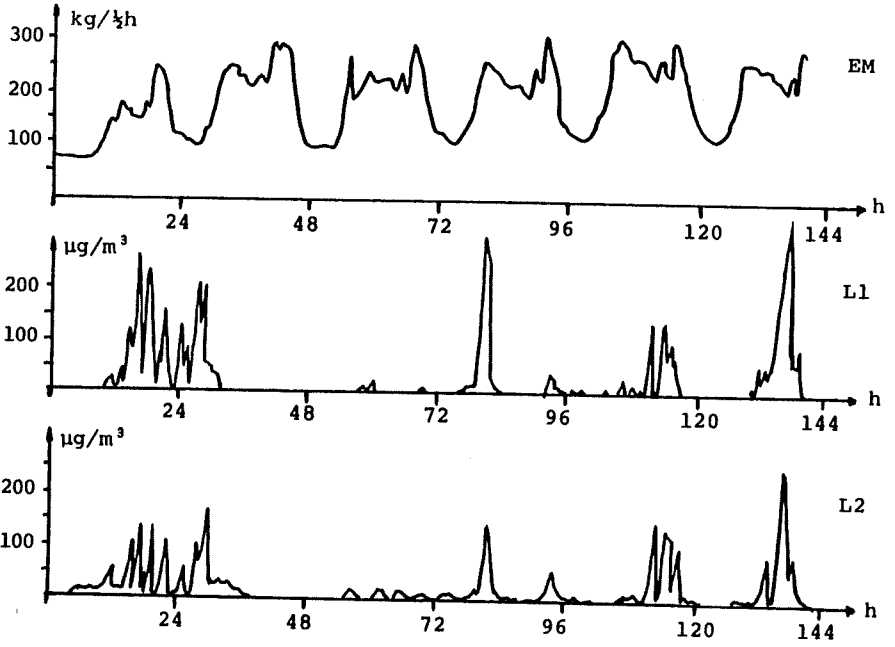
Kilden og de to målelokaliteter ligger på en ret linie i hovedvindretningen, således at røgen fra kilden oftest blæses med retning ind over målestationerne. Værdierne af $x(t)$ henholdsvis $y(t)$ og $z(t)$ er halvtimesgennemsnit af emissionen fra kilden henholdsvis immissionen på målelokaliteterne. Der foreligger 1440 samtidige målinger af de tre tidsrækker. Ved gennemsnitsdannelsen er de højfrekvente komponenter fjernet fra tidsrækkerne. Vi forudsætter, at rækkerne kan opfattes som realisationer af stationære processer.

Det er en udbredt opfattelse, at $y(t)$ og $z(t)$ afhænger (lineært?) af $x(t)$, og det primære mål for de videre analyser her er at undersøge, om en sådan eventuel påvirkning kan eftervises i korrelationsfunktioner og spektre for de tre tidsrækker.

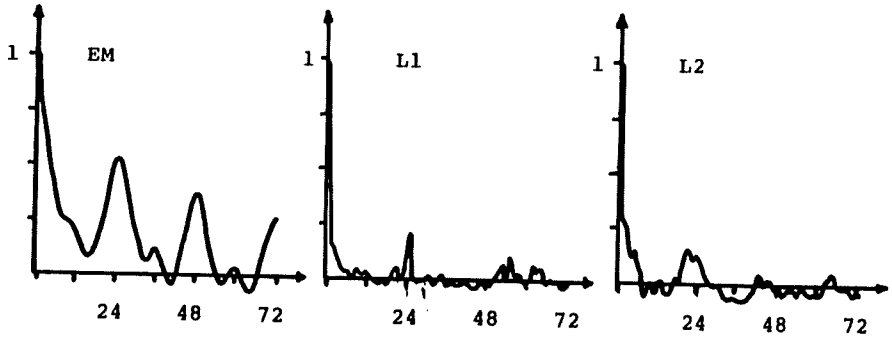
I nedenstående figur er vist et udsnit af de tre tidsrækker. Det ses, at der er en tydelig 24 h periode for $x(t)$ og noget mere uregelmæssige forhold for de to andre. Endvidere bemærker vi det store antal nuller, der er observeret i de to sidste rækker. En første visuel vurdering af disse rækker tyder ikke på en umiddelbar sammenhæng.

Som en indledende beskrivelse af disse tidsrækker er deres autokorrelationer bestemt for lags op til godt 8 døgn. I den følgende graf er de anført for lags op til 3 døgn, d.v.s. 72 tidsenheder.

Autokovarianserne og -korrelationerne er beregnet på sædvanlig måde, dog gælder, at alle figurer er baseret på hver anden af de angivne 1440 data, d.v.s. på 720 data. Dette er gjort af beregningstekniske grunde. De fundne korrelationer og spektre afviger kun uvæsentligt fra dem, der er beregnet ved anvendelse af samtlige data.



Udsnit af rækkerne x(t), y(t) og z(t).



Autokorrelationsfunktioner.

24 h perioden for x(t) genfindes tydeligt i $\hat{\rho}_{xx}(u)$. Det fremgår endvidere, at der i x(t) indgår en langsom svingning (korrelationen bliver først negativ ved et lag på 40 h). Korrelationsfunktionerne $\hat{\rho}_{yy}(u)$ og $\hat{\rho}_{zz}(u)$ afviger væsentligt fra $\hat{\rho}_{xx}(u)$. Der synes ikke at være væsentlige periodiciteter i disse sidste, og de fluktuerer omkring 0 for lags større end 6-8 h.

Da imidlertid - ifølge Tukey (1966) - "the tragic accident that killed H.R. Seiwell and his family destroyed the only man who could usefully look at a plot of autocorrelation against lag --", går vi umiddelbart videre til (power-)spektrene for autokorrelationerne. De er bestemt som cosinustransformationer af autokorrelationsfunktionerne, i.e.

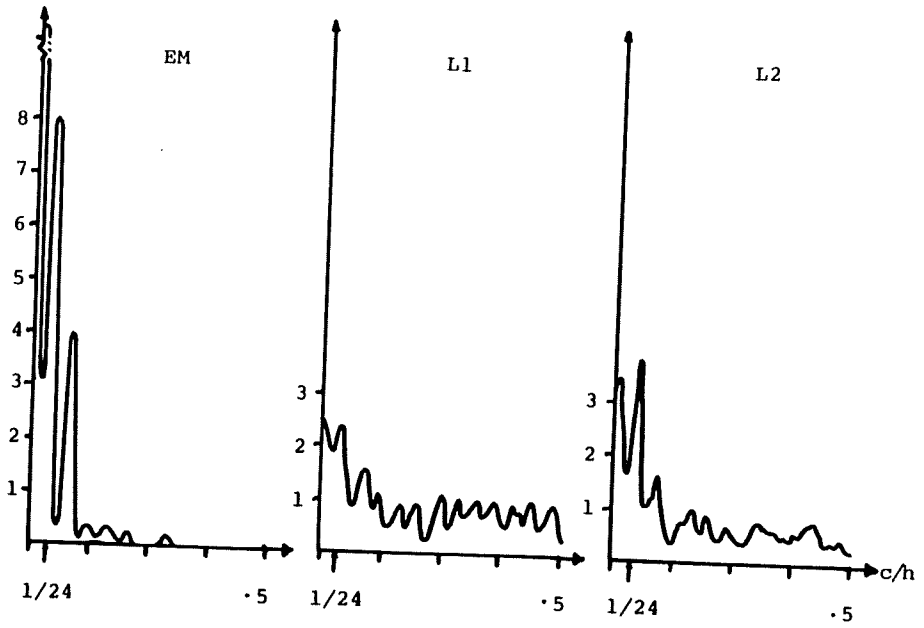
$$C(f) = \Delta[c_o + 2 \sum_{k=1}^{M-1} c_k \cos(2\pi fk\Delta)] ,$$

og dernæst udglattet ved at tage et glidende gennemsnit med vægtene 0.23, 0.54 og 0.23, d.v.s.

$$\overline{C}(f) = 0.23\ C(f-1) + 0.54\ C(f) + 0.23\ C(f+1) .$$

Dette svarer til en Hamming-udglatning. I ovenstående er Δ selvsagt tiden mellem observationer, og M er det antal led i autokorrelationsfunktionen, der medtages i beregningen. Spektrene er bestemt for en række forskellige værdier af M. Den i figuren p. 9.147 anvendte værdi af M er valgt, da den gav de generelt mest informative billeder (af de afprøvede). På trods af dette virker spektret for y(t) noget ustabilt.

Med en afstand mellem målinger på 1 h bestemmes spektrene for frekvenser mindre end ½ cycle/hour (svarende til perioden 2 h), og da det maximale lag i autokovariansfunktionen er 72 h, er spektrene bestemt for frekvenser med en afstand på $1/2 \cdot 72$ cycle/hour = 0.007 c/h.



(Power-)spektre.

For at bevare den visuelt tilgængelige egenskab, at integralet af spektret over et givet frekvensområde angiver den brøkdel af den totale varians, som de pågældende frekvenser forklarer, er det valgt at anføre spektrene i lineær skala, skønt det for mange formål (vurdering af usikkerheder på bestemmelsen af spektrene o.l.) ville være mere hensigtsmæssigt at afbilde dem i logaritmisk skala.

Det fremgår, at den væsentligste del af variansen er koncentreret i de lave frekvenser med tydelige spidser i 1/24 h^{-1} og 1/12 h^{-1}. For emissionens vedkommende er der yderligere så at sige ingen varians for frekvenser større end 1/12 h^{-1}, hvorimod såvel y(t) som z(t) har halvdelen af deres varians koncentreret i frekvenser større end 1/12 h^{-1}.

Da relationen mellem spektrene for input og output fra et lineært system er

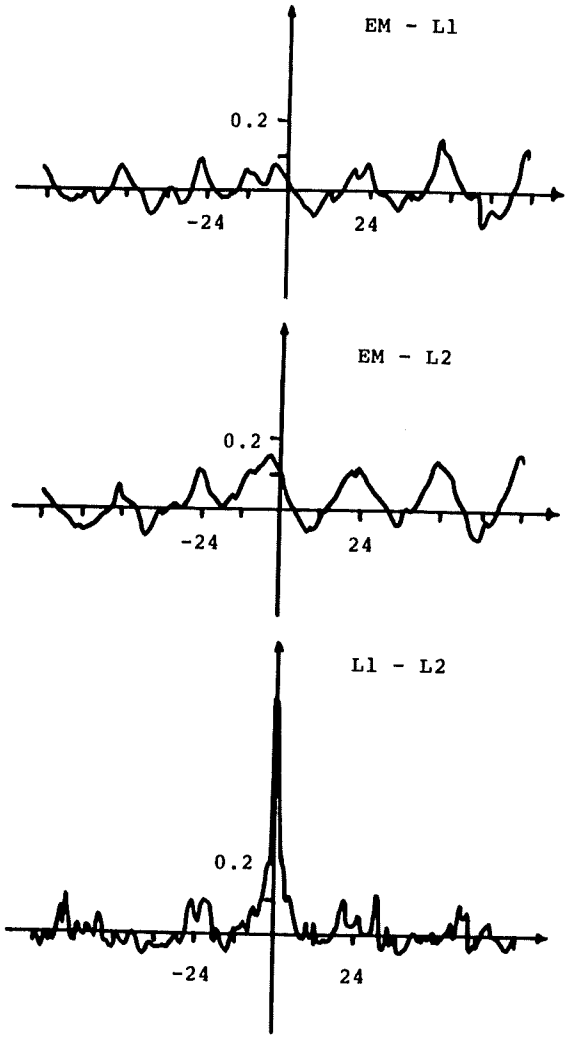
$$\Gamma_{output}(f) = Gain^2(f)\ \Gamma_{input}(f) + \Gamma_{støj}(f)\ ,$$

ses allerede af de anførte empiriske spektre, at det ikke synes rimeligt at opfatte y(t) eller z(t) som output svarende til input x(t).

Krydskorrelationsfunktionerne $\hat{\rho}_{xy}$ og $\hat{\rho}_{xz}$ udviser en tydelig periodicitet med perioden 24 h, dog en anelse faseforskudt (maximalværdien omkring 0 indtræffer for et lag på ca. -3 h). Dette er sådan set bemærkelsesværdigt, da det kunne antyde en positiv samvariation mellem målingerne af immissionen på en lokalitet og emissionen 3 h efter. Dette er selvfølgelig meningsløst rent fysisk. Forklaringen må derfor snarere være den, at variationen over døgnet af de tre rækker ikke er knyttet sammen, men er styret af andre faktorer (for x(t)'s vedkommende af energibehovet og for y(t)'s og z(t)'s vedkommende af e.g. meteorologiske forhold). Grunden til, at krydskovarianserne også for store lag er af samme størrelsesorden, skal bl.a. søges i, at der er en udpræget døgncyklus i alle tre rækker, således at målinger i to rækker med en tidsafstand på et multiplum af 24 h vil samvariere (i.e. være af samme størrelsesorden relativt til gennemsnittet). Det spiller formentlig også en rolle, at estimaterne for krydskorrelationerne er korrelerede. Vi bemærker dog, at de absolutte værdier for krydskorrelationerne er relativt små, ca. 0.1 og derunder.

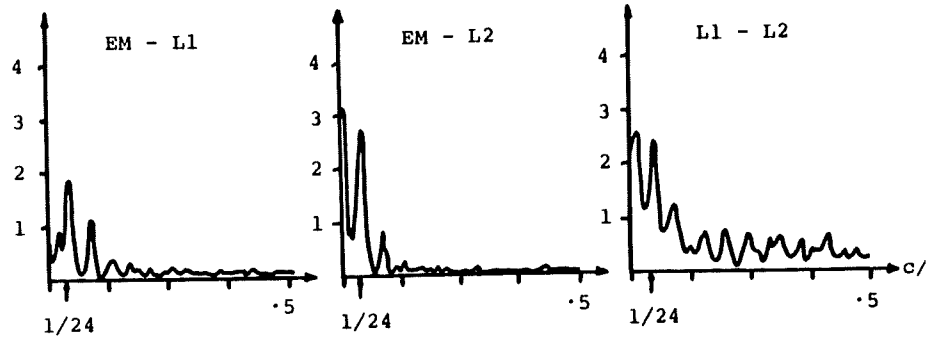
Noget andet gør sig gældende for krydskorrelationerne mellem de to immissionsrækker. Der ses at være en relativt høj korrelation for lag 0 (ca. 0.64), men derefter et ret brat fald. For lags større end 3-4 h er korrelationerne af samme størrelsesorden som de øvrige - 0.1 og derunder. Dette kunne muligvis tolkes derhen, at måleværdierne på de to stationer er styret af en eller anden ydre faktor (meteorologien), men at der ikke er nogen direkte årsagssammenhæng mellem målinger taget på de to stationer. Hvis målingerne z(t) var afhængige af værdierne af y(t), ville det med de aktuelle vindhastigheder vare mellem ca. 1/4 og ca. 1 h for en partikel at tilbagelægge afstanden mellem de to stationer. Det må dog bemærkes, at korrelationen falder en anelse hurtigere mod 0 for negative end for positive lags - noget der kunne indi-

cere en sammenhæng. På den anden side skal præciseres, at selv om målinger med tidsafstand ½ h inddrages, har krydskorrelationen dog sit maximum i 0.



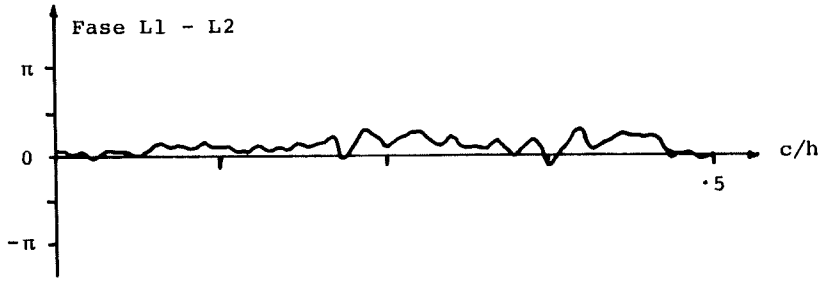
Krydskorrelationsfunktioner.

Ser vi på amplituderne for krydsspektrene, genspejles de tidligere erkendte forhold. Væsentlige dele af krydskorrelationerne ses



Krydsamplitudespektre.

at skyldes svingninger med frekvenser på 1/24 h. Dette er mest udpræget for korrelationen mellem emissionen og immissionerne på de to stationer og mindre mellem målinger på de to stationer.

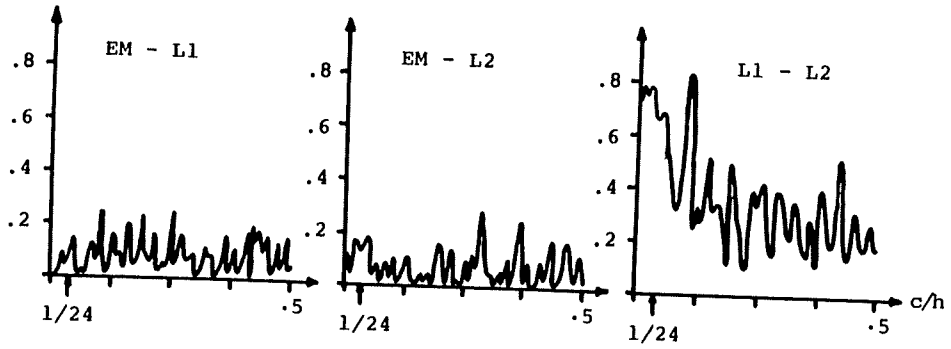


Fasespektrum for de to stationer.

Fasespektrene mellem emissionen og immissionerne på de to stationer er meget stærkt fluktuerende, hvorimod spektret for samvariationen mellem de to stationer udviser et ganske stabilt forløb.

Af de kvadrerede koherenser ses tilsvarende, at der er meget ringe sammenhæng mellem x(t) på den ene side og y(t) og z(t) på den anden side. Derimod er de relativt store mellem y(t) og z(t) -

helt op mod 0.8. (Vi må her gøre opmærksom på, at y(t) og z(t) ikke er blevet aligned med henblik på at flytte krydskorrelationsfunktionernes maximum (for $\hat{\rho}_{xy}$ og $\hat{\rho}_{xz}$) hen i 0. Dette kan give en vis bias på de første rækker).



Kvadrerede koherensspektre.

Vi bemærker også, at de lave koherenser for x(t)-y(t) og x(t)-z(t) i nogen grad kan forklare de meget uregelmæssige faseestimater, jvf. resultatet om $V(\overline{F}_{12}(f))$ p. 9.145.

For fuldstændighedens skyld skal nævnes, at analyser på såvel høj-pas som lav-pas filtrerede data ikke har afsløret tydeligere sammenhænge.

Som en præliminær konklusion af de hidtidige undersøgelser af auto- og krydskorrelationerne samt de tilsvarende spektre mellem emissionsrækken x(t) og immissionsrækkerne y(t) og z(t) synes disse ikke at indicere nogen (lineær) sammenhæng mellem emission fra punktkilden og immissionen på de to lokaliteter. Der er derimod en udpræget sammenhæng mellem y(t)- og z(t)-rækkerne til samme tidspunkter. Disse forhold indicerer, at det mere er globale forhold, der styrer immissionerne på de to målesteder, end det er emissionen fra den anførte punktkilde!

9.7 Box-Jenkins' metode

I dette afsnit skal vi kort omtale, hvorledes man kan bestemme forudsigelser af udfald af en tidsrække ved hjælp af de i afsnit 9.1.2 indførte ARMA-processer. Metoden kan sådan set ikke tilskrives Box & Jenkins alene; men da den første større samlede fremstilling af metoder og principper er Box & Jenkins (1970), er det almindeligt at anvende termen Box-Jenkins' metode.

Denne fremstilling her kan ikke gøres udtømmende, og den læser, der er interesseret i dybere resultater, må henvises til litteraturen.

Vi indleder med et afsnit, hvor vi betragter en udvidelse af klassen af ARMA-processer til visse ikke-stationære processer, de såkaldte ARIMA-processer.

9.7.1 ARIMA-processer

Mange i praksis forekommende processer udviser ikke-stationære træk. Der er måske et trend til stede eller en eller anden cyklisk bevægelse o.l. Hvis ikke-stationariteten i det væsentlige er en forstyrrelse af "middelværdiforløbet", kan man - som vi også har set tidligere - fjerne denne ved en passende differensdannelse. Hvis den "differensede" proces er stationær og kan beskrives ved en ARMA-proces, siger vi, at den oprindelige proces er en integreret (eller summeret) autoregressiv glidende gennemsnits proces eller en ARIMA-proces.

Ofte vil man først efter yderligere transformationer (e.g. logaritmering o.l.) med rimelighed kunne beskrive en ikke-stationær proces ved ARIMA-processer, men disse forhold skal vi ikke komme ind på her.

Lad ϕ være et p'te grads polynomium og θ et q'te grads, begge med rødder uden for enhedscirklen. Vi sætter

$$\varphi(B) = \phi(B)(1-B)^d = \phi(B)\nabla^d \ .$$

Dette polynomium har d rødder liggende på enhedscirklen og resten udenfor.

Definition 9.17 Vi siger med ovenstående notation, at en proces Z_t er en integreret (eller summeret) autoregressiv glidende gennemsnitsproces eller kort en ARIMA (p,d,g) proces, hvis

$$\varphi(B)\tilde{Z}_t = \theta(B)A_t \ , \tag{17}$$

eller

$$\phi(B)\nabla^d\tilde{Z}_t = \theta(B)A_t \ .$$

Da $\nabla^d\tilde{Z}_t = \nabla^d Z_t$ $(d \geq 1)$, kan vi også skrive

$$\varphi(B)Z_t = \phi(B)\nabla^d Z_t = \theta(B)A_t \ , \tag{18}$$

eller

$$\phi(B)W_t = \theta(B)A_t \ ,$$

hvor

$$W_t = \nabla^d Z_t \ .$$

Vi har med andre ord, at for en ARIMA (p,d,q)-proces vil den d'te differens $\nabla^d Z_t = W_t$ følge en stationær, invertibel ARMA (p,q)-proces.

Da

$$\nabla^d Z_t = W_t \quad \Rightarrow \quad Z_t = S^d W_t$$

kan Z_t-processen fås ved d summationer (= "integrationer") af en ARMA (p,q)-proces W_t. Derfor betegnelsen ARIMA (p,d,q)-proces.

Ved forskellige anvendelser af ARIMA-processerne får vi brug for forskellige fremstillingsformer. For lettere at kunne referere til dem, samler vi de næsten trivielle resultater i et par sætninger.

Sætning 9.42 (Differensligningsformen). Sættes

$$\varphi(B) = \phi(B)(1-B)^d = 1 - \varphi_1 B - \varphi_2 B^2 - \cdots - \varphi_{p+d} B^{p+d}$$

kan ARIMA (p,d,q)-processen (18) skrives på formen

$$Z_t = \varphi_1 Z_{t-1} + \cdots + \varphi_{p+d} Z_{t-p-d} - \theta_1 A_{t-1} - \cdots - \theta_q A_{t-q} + A_t \qquad (19)$$

Bevis Trivielt.

Sætning 9.43 (Random shock form). ARIMA-processen (18) kan skrives på formen

$$Z_t = \psi(B) A_t = A_t + \psi_1 A_{t-1} + \psi_2 A_{t-2} + \cdots , \qquad (20)$$

hvor ψ'erne kan fås ved rekursiv løsning af ligningssystemet

$$\begin{aligned}
\psi_j &= 0 , \quad j < 0 \\
\psi_0 &= 1 \\
\psi_1 &= \varphi_1 - \theta_1 \\
\psi_2 &= \varphi_1\psi_1 + \varphi_2 - \theta_2 \\
&\vdots \\
\psi_j &= \begin{cases} \varphi_1\psi_{j-1} + \cdots + \varphi_{j-1}\psi_1 + \varphi_j - \theta_j , & j \le p+d \\ \varphi_1\psi_{j-1} + \cdots + \varphi_{p+d}\psi_{j-p-d} - \theta_j , & j > p+d \end{cases}
\end{aligned}$$

Her er $\theta_j = 0$ for $j > q$.

Bevis Af (20) og (18) fås

$$\varphi(B) Z_t = \varphi(B)\psi(B) A_t = \theta(B) A_t ,$$

d.v.s.

$$\varphi(B)\psi(B) = \theta(B) \ .$$

Resultatet følger nu umiddelbart af sætning 9.7.

Q.E.D.

Sætning 9.44 (Invers form). ARIMA-processen (18) kan skrives på formen

$$A_t = \pi(B)Z_t = (1 - \sum_{j=1}^{\infty} \pi_j B^j)Z_t = Z_t - \pi_1 Z_{t-1} - \pi_2 Z_{t-2} - \cdots ,$$

eller

$$Z_t = \pi_1 Z_{t-1} + \pi_2 Z_{t-2} + \cdots + A_t \ , \qquad (21)$$

hvor π'erne bestemmes ved rekursiv løsning af ligningssystemet

$$\begin{aligned}
\pi_1 &= \varphi_1 - \theta_1 \\
\pi_2 &= \varphi_2 + \theta_1\pi_1 - \theta_2 \\
&\vdots \\
\pi_j &= \begin{cases} \varphi_j + \theta_1\pi_{j-1} + \cdots + \theta_{j-1}\pi_1 - \theta_j \ , & j \le q \\ \varphi_j + \theta_1\pi_{j-1} + \cdots + \theta_q\pi_{j-q} \ , & j > q \end{cases}
\end{aligned}$$

Her er $\varphi_j = 0$ for $j > p+d$.

Bevis Helt analogt til beviset for foregående sætning følger sætningen ved at identificere koefficienter i

$$\varphi(B) = \theta(B)\pi(B)$$

Q.E.D.

Corollar For $d \ge 1$ er

$$\sum_{j=1}^{\infty} \pi_j = 1 \ ,$$

d.v.s. den inverse fremstilling bliver af formen

$$Z_t = \overline{Z}_{t-1} + A_t \ ,$$

hvor $\overline{Z}_{t-1}$ er et vejet gennemsnit af fortidige værdier.

Bevis Da $\varphi(1) = 0$ for $d \geq 1$, og da $\theta(1) \neq 0$, er $\pi(1) = 0$, hvilket netop er det ønskede resultat.

Q.E.D.

9.7.2 Sæsonmodellen

I forbindelse med analysen af periodiske rækker vil det meget ofte være hensigtsmæssigt at arbejde med en såkaldt multiplikativ model. Man vil med en sådan model ofte kunne beskrive en forelagt tidsrække med væsentligt færre parametre, end hvis man valgte en almindelig ARIMA(p,d,q)-model.

Beskæftiger man sig f.eks. med månedlige salgstal, vil observationer fra f.eks. april måned det ene år nok minde meget om salgstallene for samme måned det foregående år, og tidsrækken vil indeholde en tydelig sæsonvariation. Hvis man da vil eliminere ikke-stationariteter og beskrive rækken, ledes man til at betragte differenser mellem målinger taget med 12 måneders mellemrum og betragte modeller, som primært lægger vægt på f.eks. den umiddelbart foregående værdi og værdien 12 måneder tidligere.

For at beskrive ovenstående løse betragtninger mere præcist, må vi først indføre en ny differensoperator.

Definition 9.18 Ved s-sæson-differensoperatoren forstås operatoren

$$\nabla_s = 1 - B^s \ ,$$

der er givet ved

$$\nabla_s x_t = x_t - x_{t-s} \ .$$

Vi indfører da den generelle multiplikative model i

Definition 9.19 Lad ϕ og θ være henholdsvis p'te og q'te grads polynomier og Φ og Θ P'te og Q'te grads polynomier, der tilfredsstiller stationaritets- og invertibilitetskrav. Vi siger da, at processen Z_t følger en multiplikativ $(p,d,q) \times (P,D,Q)_s$ sæsonmodel med residualproces A_t, såfremt

$$\phi(B)\Phi(B^s)\nabla^d \nabla_s^D Z_t = \theta(B)\Theta(B^s)A_t \ .$$

Vi anfører et lille eksempel til at klargøre begrebet.

Eksempel 40 Vi betragter en $(2,1,1) \times (1,0,0)_{12}$ model med definerende polynomier ϕ, Φ og θ, d.v.s.

$$\phi(B)\Phi(B^{12})\nabla Z_t = \theta(B)A_t \ ,$$

eller

$$(1-\phi_1 B-\phi_2 B^2)(1-\Phi_1 B^{12})(1-B)Z_t = (1-\theta_1 B)A_t \ ,$$

d.v.s.

$$\begin{aligned}[1-(1+\phi_1)B-(\phi_1-\phi_2)B^2+\phi_2 B^3-\Phi_1 B^{12} \\ +\Phi_1(1+\phi_1)B^{13}-\Phi_1(\phi_1-\phi_2)B^{14}-\phi_2\Phi_1 B^{15}]Z_t = (1-\theta_1 B)A_t \ .\end{aligned}$$

Vi ser altså, at der er tale om differenser af op til femtende orden i Z'er sat lig differenser af første orden i A'er, men hele systemet er beskrevet ved kun 4 parametre. Det er heri begrebets styrke ligger.

□

Som ovenstående antyder, er der ingen principielle forskelle mellem en almindelig ARIMA-model og en multiplikativ sæsonmodel, og vi vil derfor ikke i det følgende specielt anføre resultater for sæsonmodellerne.

9.7.3 Forudsigelser i ARIMA-processer

Vi vil nu søge en forudsigelse af udfaldet af en ARIMA-proces om ℓ tidsenheder. Forudsigelsen skal baseres på nuværende og tidligere observationer. For at kunne opstille en sådan forudsigelse må vi indføre et optimalitetskriterium, og vi vælger at søge at minimalisere den forventede kvadratiske fejl på forudsigelsen, d.v.s. vi søger at minimalisere størrelsen

$$E[(Z_{t+\ell} - \hat{Z}_t(\ell))^2] ,$$

hvor $\hat{Z}_t(\ell)$ er vor forudsigelse til tidspunkt t af, hvad der vil ske til tidspunkt $t+\ell$, d.v.s. til et tidspunkt ℓ tidsenheder senere.

Vi formulerer løsningen i

Sætning 9.45 Vi betragter ARIMA-processen (18), og vi sætter

$$\hat{Z}_t(\ell) = \psi_\ell A_t + \psi_{\ell+1} A_{t-1} + \psi_{\ell+2} A_{t-2} + \cdots .$$

Da er $\hat{Z}_t(\ell)$ den forudsigelse, der har den mindste forventede kvadratiske fejl.

Bevis Lad

$$\hat{Z}_t(\ell) = \psi^*_\ell A_t + \psi^*_{\ell+1} A_{t-1} + \psi^*_{\ell+2} A_{t-2} + \cdots$$

Vi har nu, idet vi benytter (20),

$$Z_{t+\ell} = A_{t+\ell} + \psi_1 A_{t+\ell-1} + \cdots + \psi_{\ell-1} A_{t+1} + \psi_\ell A_t + \psi_{\ell+1} A_{t-1} + \cdots ,$$

hvorfor

$$Z_{t+\ell} - \hat{Z}_t(\ell) = A_{t+\ell} + \cdots + \psi_{\ell-1} A_{t+1} + \sum_{j=0}^{\infty} (\psi_{\ell+j} - \psi^*_{\ell+j}) A_{t-j} \ .$$

Da A_τ'erne er indbyrdes ukorrelerede, bliver

$$E[(Z_{t+\ell} - \hat{Z}_t(\ell))^2] = (1 + \psi_1^2 + \cdots + \psi_{\ell-1}^2)\sigma_a^2 + \sum_{j=0}^{\infty} (\psi_{\ell+j} - \psi^*_{\ell+j})^2 \sigma_a^2,$$

og heraf følger sætningen, idet den første parentes er uafhængig af ψ^*, og det sidste led er netop lig 0 for $\psi^*_{\ell+j} = \psi_{\ell+j}$, $j = 0,1,2 \cdots$.

Q.E.D.

Corollar 1 Med notationen fra foregående sætning haves

$$\hat{Z}_t(\ell) = E(Z_{t+\ell} | Z_t, Z_{t-1}, \cdots) \ ,$$

d.v.s. den bedste forudsigelse er lig den betingede middelværdi af $Z_{t+\ell}$ givet alle målinger inden tid t.

Bevis Vi har

$$Z_{t+\ell} = A_{t+\ell} + \cdots + \psi_{\ell-1} A_{t+1} + \hat{Z}_t(\ell) \ .$$

Den betingede middelværdi af et random shock er

$$E(A_\tau | Z_t, Z_{t-1}, \cdots) = \begin{cases} A_\tau & \tau \leq t \\ 0 & \tau > t \end{cases} \ .$$

Ved benyttelse af dette følger resultatet.

Q.E.D.

Ved indførelse af en særlig notation for den betingede forventningsværdi kan de tidligere resultater udtrykkes mere operationelt. Vi anfører derfor

Definition 9.20 For den betingede middelværdi af en stokastisk variabel Y givet alle observationer af processen Z_T taget før tidspunkt t anvendes betegnelsen

$$[Y] = [Y]_t = E(Y|Z_t, Z_{t-1}, \cdots) ,$$

hvor index t kun anføres, hvis tvivl er mulig.

Med denne notation har vi så

Sætning 9.46 (Forudsigelsesligning). For ARIMA-processen (18) har vi følgende udtryk til bestemmelse af den bedste forudsigelse ℓ tidsenheder frem:

$$\hat{Z}_t(\ell) = [Z_{t+\ell}] = \varphi_1[Z_{t+\ell-1}] + \cdots + \varphi_{p+d}[Z_{t+\ell-p-d}] - \theta_1[A_{t+\ell-1}] - \cdots - \theta_q[A_{t+\ell-q}] + [A_{t+\ell}] ,$$

hvor

$$[Z_{t-j}] = Z_{t-j} , \quad j = 0,1,2,\cdots$$
$$[Z_{t+j}] = \hat{Z}_t(j) , \quad j = 1,2,\cdots$$
$$[A_{t-j}] = A_{t-j} = Z_{t-j} - \hat{Z}_{t-j-1}(1) , \quad j = 0,1,2,\cdots$$
$$[A_{t+j}] = 0 , \quad j = 1,2,\cdots$$

Bevis Resultatet er en triviel følge af det tidligere. Bemærkes skal dog, at

$$\hat{Z}_{t-j-1}(1) = \psi_1 A_{t-j-1} + \psi_2 A_{t-j-2} + \cdots = Z_{t-j} - A_{t-j} .$$

Q.E.D.

Bemærkning Sætningen viser, hvorledes en forudsigelse ℓ tidsenheder frem opbygges ved hjælp af kortere forudsigelser. Ved en praktisk beregning findes forudsigelserne derfor rekursivt. Hvis der er et moving average led i modellen, får vi brug for at beregne de ukendte A'er. Dette gøres som sagt ved hjælp af udtrykket $Z_{t-j} - \hat{Z}_{t-j-1}(1)$, men dette kræver kendskab til tid-

ligere værdier af A'er etc. Man starter derfor som regel forudsigelsesprocessen et stykke tilbage i tiden og sætter A'ernes initialværdier lig deres forventningsværdier, nemlig 0.

Eksempel 9.41 Vi betragter processen

$$Z_t = \frac{5}{6} Z_{t-1} - \frac{1}{6} Z_{t-2} + A_t - \frac{1}{4} A_{t-1} .$$

Vi vil først undersøge, hvilken proces der er tale om. Vi opskriver processen på formen

$$(1 - \frac{5}{6} B + \frac{1}{6} B^2) Z_t = (1 - \frac{1}{4} B) A_t$$

eller

$$\frac{1}{6} (B-2)(B-3) Z_t = \frac{1}{4} (4-B) A_t .$$

Der er altså tale om en ARMA (2,1) model, og den ses at være såvel stationær som invertibel.

Vi skal dernæst bestemme forudsigelsesligningerne. Vi har

$$Z_{t+\ell} = \frac{5}{6} Z_{t+\ell-1} - \frac{1}{6} Z_{t+\ell-2} + A_{t+\ell} - \frac{1}{4} A_{t+\ell-1}$$

d.v.s

$$Z_{t+1} = \frac{5}{6} Z_t - \frac{1}{6} Z_{t-1} + A_{t+1} - \frac{1}{4} A_t ,$$

hvorfor

$$\hat{Z}_t(1) = \frac{5}{6} [Z_t] - \frac{1}{6} [Z_{t-1}] + [A_{t+1}] - \frac{1}{4} [A_t]$$

$$= \frac{5}{6} Z_t - \frac{1}{6} Z_{t-1} - \frac{1}{4} A_t .$$

Analogt

$$\hat{Z}_t(2) = \frac{5}{6}\,[Z_{t+1}] - \frac{1}{6}\,[Z_t] + [A_{t+2}] - \frac{1}{4}\,[A_{t+1}]$$

$$= \frac{5}{6}\,\hat{Z}_t(1) - \frac{1}{6}\,Z_t$$

$$\hat{Z}_t(3) = \frac{5}{6}\,\hat{Z}_t(2) - \frac{1}{6}\,\hat{Z}_t(1)$$

$$\vdots$$

$$\hat{Z}_t(\ell) = \frac{5}{6}\,\hat{Z}_t(\ell-1) - \frac{1}{6}\,\hat{Z}_t(\ell-2)\ .$$

Vi ser, at vi skal bruge A_t i $\hat{Z}_t(1)$. Ifølge sætning 9.46 findes denne størrelse som $Z_t - \hat{Z}_{t-1}(1)$. Nu er den sidste størrelse lig

$$\hat{Z}_{t-1}(1) = \frac{5}{6}\,Z_{t-1} - \frac{1}{6}\,Z_{t-2} - \frac{1}{4}\,A_{t-1}\ .$$

For at finde denne må vi kende A_{t-1} o.s.v. Man vil derfor i praksis starte forudsigelsesproceduren e.g. k tidsenheder tilbage og sætte de ukendte A'er lig 0, d.v.s. sætte

$$\hat{Z}_{t-k}(1) \simeq \frac{5}{6}\,Z_{t-k} - \frac{1}{6}\,Z_{t-k-1}\ .$$

Derfor er

$$A_{t-k+1} = Z_{t-k+1} - \hat{Z}_{t-k}(1) \simeq Z_{t-k+1} - \frac{5}{6}\,Z_{t-k} + \frac{1}{6}\,Z_{t-k-1}\ .$$

Dernæst findes

$$\hat{Z}_{t-k+1}(1) = \frac{5}{6}\,Z_{t-k+1} - \frac{1}{6}\,Z_{t-k} - \frac{1}{4}\,A_{t-k+1}\ ,$$

hvor vi indsætter den approximative værdi for A_{t-k+1}. Dette giver så

$$A_{t-k+2} = Z_{t-k+2} - \hat{Z}_{t-k+1}(1)\ ,$$

og så fremdeles.

□

Når man på basis af observationer $\cdots, Z_{t-1}, Z_t$, har beregnet forudsigelser $\hat{Z}_t(1), \hat{Z}_t(2), \cdots, \hat{Z}_t(\ell), \cdots$, kan man selvfølgelig, efter at observationen Z_{t+1} er indløbet, på helt sædvanlig vis beregne forudsigelserne $\hat{Z}_{t+1}(1), \hat{Z}_{t+1}(2)$ etc. Der findes dog en langt lettere måde at foretage denne opdatering på, nemlig

Sætning 9.47 (Opdatering af forudsigelser). Efter fremkomsten af observationen Z_{t+1} kan forudsigelserne opdateres efter formlen

$$\hat{Z}_{t+1}(\ell-1) = \hat{Z}_t(\ell) + \psi_{\ell-1}A_{t+1} ,$$

d.v.s. vort gæt over, hvad der vil ske til tiden t+ℓ, justeres ved addition af leddet $\psi_{\ell-1}A_{t+1}$.

Bevis Ifølge sætning 9.44 har vi

$$\hat{Z}_{t+1}(\ell-1) = \psi_{\ell-1}A_{t+1} + \psi_\ell A_t + \psi_{\ell+1}A_{t-1} + \cdots$$

$$\hat{Z}_t(\ell) = \psi_\ell A_t + \psi_{\ell+1}A_{t-1} + \cdots ,$$

hvorfor

$$\hat{Z}_{t+1}(\ell-1) - \hat{Z}_t(\ell) = \psi_{\ell-1}A_{t+1} .$$

Q.E.D.

Den statistiske usikkerhed på forudsigelsen fremgår af

Sætning 9.48 Variansen på forudsigelsesfejlen ℓ tidsenheder fremme er

$$\sigma^2(\ell) = E(\{Z_{t+\ell} - \hat{Z}_t(\ell)\}^2) = \sigma_a^2\{1 + \psi_1^2 + \cdots + \psi_{\ell-1}^2\} .$$

Er s_a^2 et skøn over σ_a^2, bliver et (approximativt) 1-α konfidensinterval for $Z_{t+\ell}$

$$\hat{Z}_t(\ell) \pm u_{\alpha/2}s_a\{1 + \psi_1^2 + \cdots + \psi_{\ell-1}^2\}^{0.5} .$$

Bevis Det første resultat følger umiddelbart af beviset for sætning 9.44, det sidste af nogle overvejelser om asymptotisk normalitet.

Q.E.D.

9.7.4 Identifikation af og estimation i en ARIMA-proces

Ved identifikation af en ARIMA (p,d,q)-proces forstås fastlæggelse af de i praksis næsten altid ukendte størrelser p, d og q. Estimationen er selvfølgelig estimationen af koefficienterne i polynomierne ϕ og θ (efter at p, d og q er fastlagte).

Først bestemmes graden d af differensdannelse. Her benyttes, at de stationære processer har autokorrelationer, der går mod 0 for lag nr. $\to \infty$. Man undersøger derfor den empiriske autokorrelationsfunktion for $\nabla^d z_t$ for forskellige værdier af d. Når autokorrelationen dør ud "tilstrækkelig hurtigt", har man nået den nødvendige grad af differensdannelse.

Ovenstående virker måske nok temmelig subjektivt, og det kan derfor være nyttigt at mærke sig, at d i praksis som regel er 0, 1 eller 2, og at det oftest er tilstrækkeligt at undersøge de første ca. 20 autokorrelationer i den oprindelige række og i de differensede rækker.

Hvis man arbejder med en sæsonmodel skal man selvsagt også bestemme graden D af sæsondifferensdannelse, og dette sker helt analogt til ovenstående.

Er der nu tale om en ren AR(p)- eller en ren MA(q)-proces, kan p respektive q fastlægges ved at undersøge, om den partielle respektive den almindelige autokorrelationsfunktion dør ud efter p respektive efter q lags, jfr. sætningerne 9.17, 9.19, 9.21, 9.22 og bemærkningen p. 9.53.

Ved vurderingen af korrelationsfunktionernes absolutte størrelse kan man benytte

Sætning 9.49 I en AR(p)-proces gælder, at variansen på den partielle autokorrelation

$$\hat{V}(\hat{\phi}_{kk}) \simeq \frac{1}{n}, \quad k \geq p+1,$$

hvor n er det antal observationer, $\hat{\phi}_{kk}$ er estimeret på basis af.

Sætning 9.50 I en MA(q)-proces gælder

$$\hat{V}(r_k) \simeq \frac{1}{n}\{1 + 2(r_1^2 + \cdots + r_q^2)\} \qquad k \geq q+1$$

Beviser Forbigås.

Bemærkning Det skal anføres, at de anførte skøn $\hat{\phi}_{kk}$ og r_k er approximativt normalt fordelte med middelværdier 0 og de anførte varianser. Dette kan udnyttes ved konstruktion af tilnærmede tests for hypoteser om, at funktionerne forsvinder efter et vist trin.

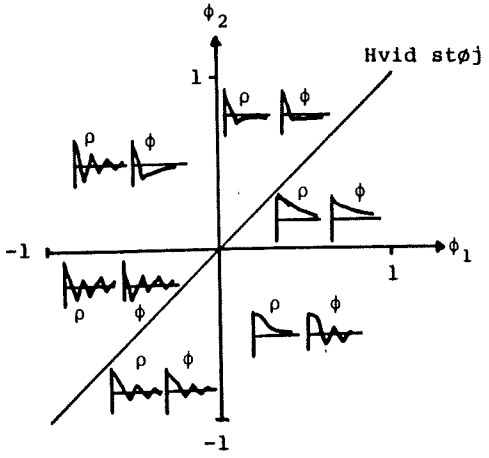
Fastlæggelse af p og q i en ARMA (p,q)-model er noget vanskeligere.

Af sætning 9.24 følger, at autokorrelationerne for lags ≥ q tilfredsstiller differensligningen

$$\rho_k = \phi_1\rho_{k-1} + \cdots + \phi_p\rho_{k-p} .$$

Man kan således lede efter et regulært mønster for k ≥ q. Af corollar 2 til samme sætning følger, at autokorrelationen går mod 0 som dæmpede eksponential- og/eller sinusfunktioner, hvis q < p. Hvis q ≥ p, vil der være q-p+1 begyndelsesværdier, der ikke tilfredsstiller dette krav.

Nedenstående figur (jvf. Box & Jenkins (1970)) kan være nyttig, idet den viser, hvor omfattende klassen af ARMA (1,1)-modeller er.



Autokorrelations- og partielle autokorrelationsfunktioner ρ_k og ϕ_{kk} for forskellige ARMA (1,1)-modeller.

Estimationsproblemet i ARIMA-processerne er ganske kompliceret. I afsnit 9.7.2 har vi anført nogle resultater, der kan danne udgangspunkt for en præliminær estimation af parametrene. Her må især nævnes sætning 9.19 om Yule-Walker ligningerne for en autoregressiv proces. Ved at indsætte empirisk bestemte korrelationer i disse og løse ligningssystemet kan man få estimater for parametrene i AR(p)-processen. For mindre værdier af q kan resultatet i sætning 9.17 benyttes ved fastlæggelse af parametrene i en glidende gennemsnitsproces.

For såvel MA(2)- som AR(2)- og ARMA(1,1)-processerne findes i Box & Jenkins (1970) kurveblade, der relaterer empiriske korrelationer til parameterskøn i de omtalte processer.

Forudsætter man, at de involverede observationer er normalt fordelte, kan man finde maximum likelihood estimater for parametrene. Det er ret kompliceret at beregne likelihoodfunktionen, og maksimaliseringen må foregå iterativt. Der findes efterhånden en række standardprogrammer, der kan udføre disse ting. Nævnes kan her en række rutiner fra IMSL-biblioteket. Vi skal nedenfor kort antyde denne teknik i et eksempel.

Eksempel 9.42 Lad os betragte en ARIMA(1,0,1)-proces givet ved

$$(1 - \phi B)Z_t = (1 - \theta B)A_t \ ,$$

hvor A_t'erne forudsættes uafhængige og $N(0,\sigma_a^2)$-fordelte, og at $E(Z_t) = 0$. Vi forudsætter, at der foreligger observationer $Z_1,\cdots,Z_n$. Hvis man også kendte Z_0 og A_0, kan man successivt beregne

$$\left.\begin{aligned} A_1 &= Z_1 - \phi Z_0 + \theta A_0 \\ A_2 &= Z_2 - \phi Z_1 + \theta A_1 \\ &\vdots \\ A_n &= Z_n - \phi Z_n + \theta A_{n-1} \end{aligned}\right\} \ . \qquad (22)$$

Frekvensfunktionen for $(A_1,\cdots,A_n)'$ er som bekendt

$$f(a_1,\cdots,a_n) = \frac{1}{\sqrt{2\pi}^n}\frac{1}{\sigma_a^n}\exp\left(-\frac{1}{2\sigma_a^2}\sum_i a_i^2\right) \ .$$

Likelihoodfunktionen for (ϕ,θ) betinget af det anførte valg af (Z_0,A_0) bliver derfor, idet vi betegner de ved ligningssystemet (22) bestemte værdier af A_i med $A_i(\phi,\theta|Z_0,A_0)$,

$$L(\phi,\theta,\sigma_a^2) = \frac{1}{\sqrt{2\pi}^n}\frac{1}{\sigma_a^n}\exp\left(-\frac{1}{2\sigma_a^2}\sum_i A_i^2(\phi,\theta|Z_0,A_0)\right) \ .$$

Den værdi af (ϕ,θ,σ_a^2), der maksimaliserer denne funktion (eller dens logaritme), er da betinget maksimum likelihood estimat for processens parametre.

Problemet er nu at finde Z_0 og A_0. En metode er at bruge den første del af rækken til at estimere A_t'erne og kun bruge den sidste del i estimationen. En anden måde er at finde "back forecasts" af de tidligere værdier. Ideen i dette er at "vende" tidsaksen og "forudsige" de "fortidige" værdier ved hjælp af de sædvanlige differensligninger. Begrundelserne for

denne metodik skal vi ikke komme nærmere ind på her, men blot henvise den interesserede læser til Box & Jenkins (1970).

□

9.7.5 Et eksempel

Vi skal i dette afsnit give et eksempel på en foreløbig anvendelse af ARIMA-processer til forudsigelser (se Conradsen et al. 1977).

Problemet vedrører prædiktion af extreme vandstande ved Esbjerg, især i stormvejrsperioder. Der findes en hydrodynamisk-numerisk grundmodel for Nordsøen (udviklet på Meteorologisk Institut), den såkaldte HN-model, der kan bruges hertil; men denne model giver ikke uvæsentlige forudsigelsesfejl, eller, som disse fejl også kaldes, residualer. Vi gennemgår nedenfor en analyse med det sigte at forudsige HN-modellens residualer. Residualet til tiden t er defineret ved

$$z_t = E_t - HN_t ,$$

hvor E_t er den observerede vandstand ved Esbjerg til tiden t, og HN_t er HN-modellens beregnede vandstand hørende til tiden t. Der tilstræbes herved en direkte prognosticering af forskellen mellem HN-modellens beregnede vandstande og de faktisk forekommende vandstande uden inddragelse af yderligere observationsrækker.

Til denne undersøgelse er benyttet data fra to tidsperioder; nemlig en lang, rolig periode uden storme, bestående af oktober og november 1976, og en kortere periode fra februar 1967, indeholdende en kraftig storm. Metoden er en undersøgelse af, om residualerne kan beskrives ved hjælp af ARIMA-processer.

Det spørgsmål, som er interessant at få belyst, er, om residualernes variation er rent tilfældig, eller om de også indeholder systematisk variation, d.v.s. en vis grad af hukommelse. Dette kan undersøges ved at betragte residualernes autokorrelationsfunktion. Er residualerne autokorrelerede, vil det være muligt ud fra observerede værdier af HN-modellens residualer at

udtale sig om de fremtidige residualers størrelse, hvorved HN-modellens fremtidige beregnede vandstande kan korrigeres til at give bedre forudsigelser.

Vi betragter residualernes autokorrelationsfunktioner for hver af de to perioder. Disse er vist på figuren p. 9.170. Antallet af observationer for perioden oktober/november 1976 er 1464, og for perioden februar 1967 er det 216.

Det ses, at autokorrelationsfunktionerne kun langsomt går mod nul, og at der optræder en "sæson"-effekt på ca. 12 timer, svarende til en tidevandsperiode. Denne sæsoneffekt er til stede ved begge perioder, men er mest fremtrædende i perioden med roligt vejr. Dette kan skyldes, at tidevandsperioden forstyrres under kraftig storm. For at tage hensyn til sæsoneffekten indføres en sæsonparameter. Problemet søges løst ved en transformation af formen

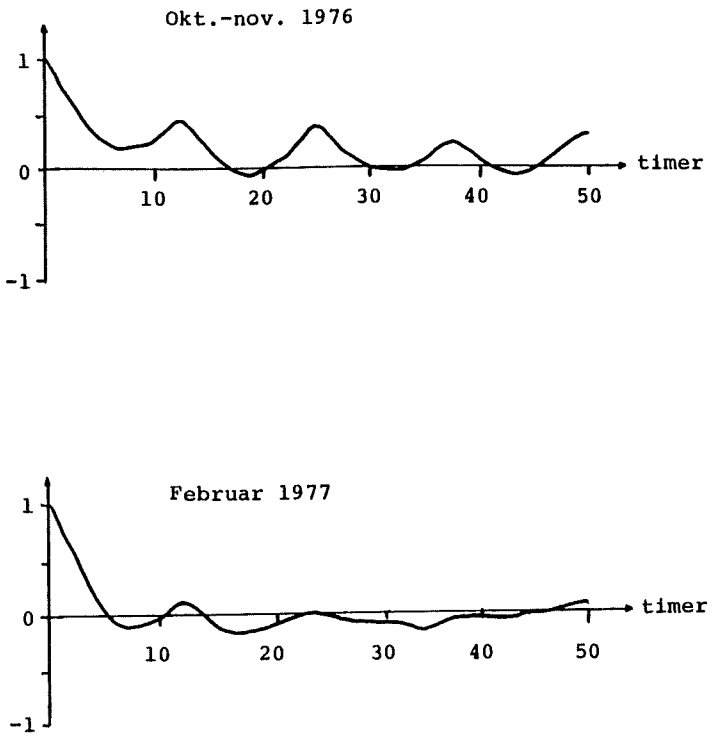
$$y_t = z_t - \alpha z_{t-12} ,$$

hvor konstanten α kan estimeres. Her er $z_t = E_t - HN_t$ residualet til tiden t.

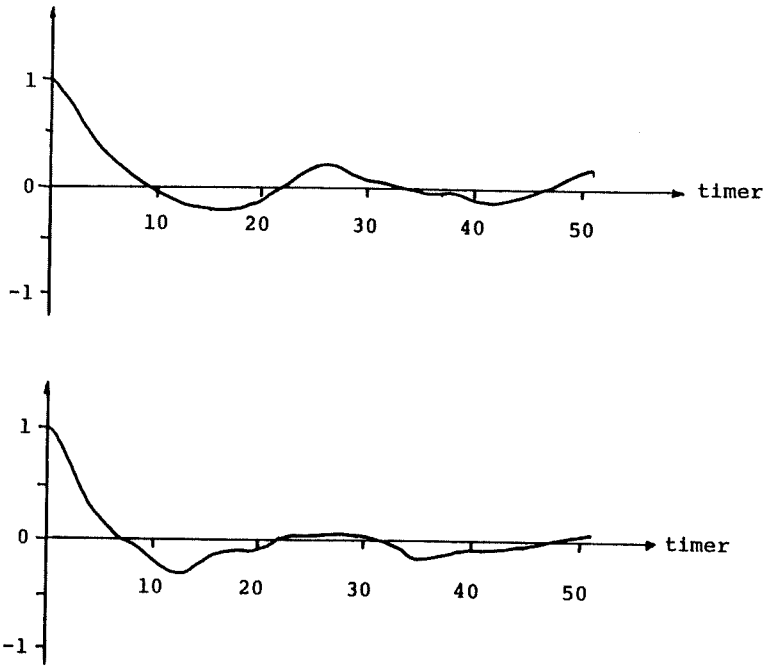
Rimeligheden af transformationen er endnu ikke undersøgt i detaljer, ligesom der kun er udført et foreløbigt skøn over α. En umiddelbar vurdering af autokorrelationsfunktionen antyder, at $\alpha = \frac{1}{2}$ er et rimeligt første gæt, d.v.s.

$$y_t = z_t - \frac{1}{2} z_{t-12} .$$

På figuren p. 9.171 ses autokorrelationensfunktionen for den sæsontransformerede række y_t.



Autokorrelationsfunktioner for HN-modellens residualer.

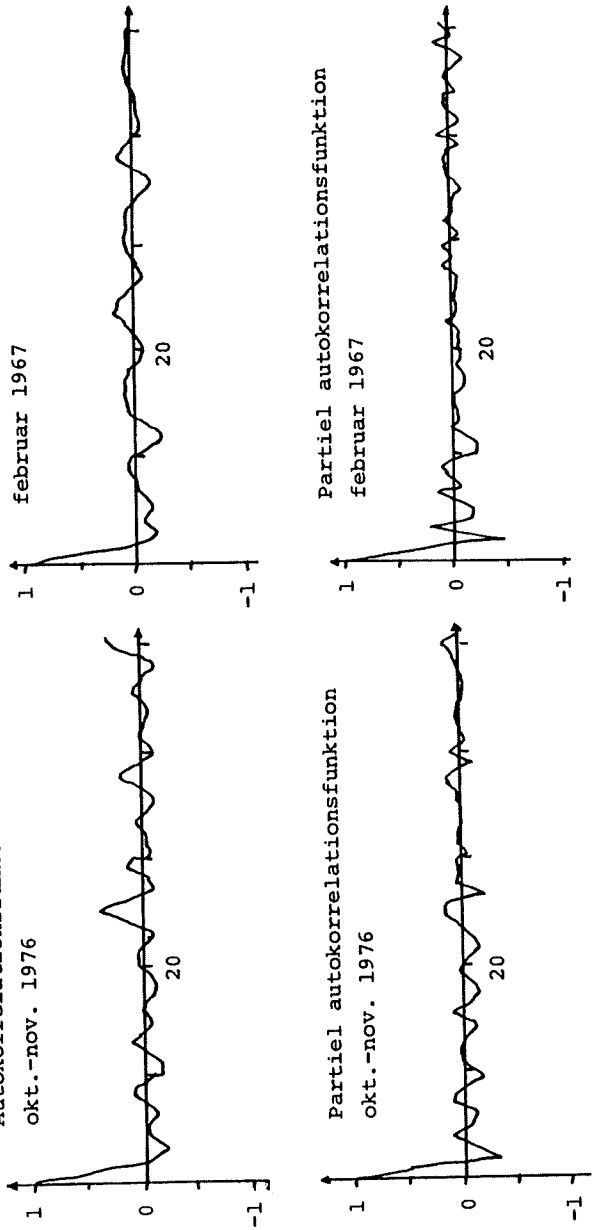


Autokorrelationer for de sæsontransformerede rækker.

Det ses, at der stadig er lidt "sæson"-effekt tilbage, så der bør foretages yderligere analyser. Til trods herfor er det dog fundet rimeligt at arbejde videre med denne transformation. Af figuren ses endvidere, at autokorrelationsfunktionen kun langsomt går mod nul, men ved at differenstransformere rækken y_t på på følgende måde

$$x_t = y_t - y_{t-1} ,$$

fås autokorrelationer og partielle autokorrelationer for rækken x_t, som vist på figuren p. 9.172.



Autokorrelations- og partielle autokorrelationsfunktioner
for de differenstransformerede rækker.

Af figuren fremgår, at for 1'ste ordens differenser går autokorrelationsfunktionen rimelig hurtigt mod nul, hvilket antyder, at den 1'ste ordens differenstransformerede række kan approximeres ved en ARMA-model. Det fremgår endvidere af figurerne, at der er en tydelig lighed mellem autokorrelations- og de partielle autokorrelationsfunktioner for de to perioder. Dette betyder, at variationen af residualerne kan beskrives ved den samme type proces for de to perioder.

Da de partielle autokorrelationsfunktioner er forsvindende, kan den "sæson"-korrigerede residualproces $y_t = z_t - \frac{1}{2} z_{t-12}$, hvor z_t er HN-modellens residualer, beskrives ved en ARIMA (2,1,0)-proces. Dette resultat er gyldigt for begge de betragtede perioder, d.v.s. både for stormvejrsperioden og den mere rolige periode. Man kan hæfte sig ved, at der ikke indgår glidende gennemsnit i denne model.

Rækken x_t kan altså skrives

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + a_t \ ,$$

hvorfor

$$y_t = \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \varphi_3 y_{t-3} + a_t \ .$$

Parametrene φ_1, φ_2 og φ_3 er for hver af de to perioder estimeret ved numerisk maskimalisering af likelihoodfunktionen. Hertil benyttedes routinen FTMAXL fra subroutinebiblioteket IMSL, der udfører denne maksimalisering iterativt. Der er fundet følgende parameterestimater:

	$\hat{\varphi}_1$	$\hat{\varphi}_2$	$\hat{\varphi}_3$
rolig periode	1.5879	- 0.9153	0.3274
stormperiode	1.7248	- 1.1963	0.4714

Det ses, at estimaterne ikke adskiller sig væsentligt fra hinanden for de to perioder; men et statistisk test viser dog, at

de næppe kan anses for at være ens. Derfor er alle de følgende analyser foretaget med begge parametersæt, og der har vist sig en fin overensstemmelse mellem de resultater, der opnås.

For at bedømme, hvor godt rækken y_t beskrives, sammenlignes rækkens oprindelige varians med ARIMA-processens residualvarians, hvilket giver følgende værdier

	Estimationsgrundlag	
	roligt vejr	stormvejr
y_t's oprindelige varians	267 cm^2	545 cm^2
ARIMA-processens residualvarians	36 cm^2	76 cm^2
Residualvarians i % af oprindelig varians	14 %	14 %

Det ses, at for begge perioder beskrives ca. 86% af den oprindelige varians, og at der på dette punkt er overensstemmelse mellem de to perioder.

Da der således ses at kunne opnås en god beskrivelse af HN-modellens fejl ved hjælp af en ARIMA-proces, er det relevant at forsøge at forudsige fremtidige fejl ved hjælp heraf. Ligningerne til bestemmelse af fremtidige y-værdier bliver generelt, idet vi udelukker ^ over φ'erne,

$$y_{t+k} = \varphi_1 y_{t+k-1} + \varphi_2 y_{t+k-2} + \varphi_3 y_{t+k-3} \cdot$$

Ligningen benyttes successivt ved at starte med en 1-times forudsigelse

$$\hat{y}_t(1) = \varphi_1 y_t + \varphi_2 y_{t-1} + \varphi_3 y_{t-2} ,$$

og derefter

$$\hat{y}_t(2) = \varphi_1 \hat{y}_t(1) + \varphi_2 y_1 + \varphi_3 y_{t-1} , \text{ osv.}$$

En forudsigelse af vandstanden ved Esbjerg 1 time frem findes da ved

$$\hat{E}_t(1) = \hat{z}_t(1) + HN_{t+1} ,$$

d.v.s.

$$\hat{E}_t(1) = \hat{y}_t(1) + 0.5(E_{t-11} - HN_{t-11}) + HN_{t+1} ,$$

hvor $\hat{y}_t(1)$ findes beskrevet som ovenfor.

Denne forudsigelsesmodel for HN-modellens fejl anvendes således til at korrigere HN-modellens forudsigelser, og denne korrektion foretages ved hver ny indkommet observeret timevandstand. Modellen er afprøvet på en stormflodsperiode. Der er her valgt en periode, som ikke har været benyttet ved estimationen af parametrene φ_1, φ_2 og φ_3. Perioden består af de 6 første døgn af januar 1976. Denne periode indeholder den voldsomste stormflod i dette århundrede.

Der er foretaget gennemregninger med begge sæt værdier af φ_1, φ_2 og φ_3. I nedenstående tabel ses varianser og spredninger af forudsigelsesfejlene ved forudsigelser af forskellig længde.

Længde af forudsigelse	Estimationsgrundlag			
	Roligt vejr		Stormvejr	
	Varians (cm^2)	Spredning (cm)	Varians (cm^2)	Spredning (cm)
1 time	153	12	148	12
2 timer	595	24	582	24
3 timer	1018	32	1006	32
4 timer	1242	35	1222	35
5 timer	1390	37	1355	37
6 timer	1552	39	1500	39

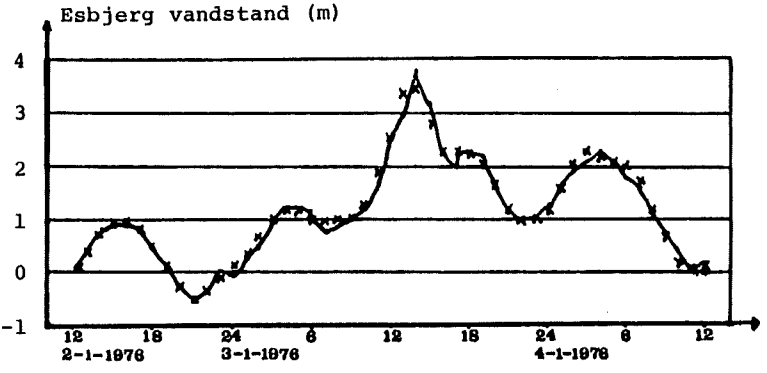
Forudsigelsesresultater for 1.-6. januar 1976.

Til sammenligning haves, at HN-modellens forudsigelsesfejl for denne periode udviser en varians på 1685 cm^2 og en spredning på 41 cm.

Man hæfter sig ved, at de to parametersæt resulterer i forudsigelser af samme godhed. De enkelte forudsigelser er i øvrigt næsten identiske.

Det ses således, at der især ved de kortere forudsigelser opnås en væsentlig nedbringelse af forudsigelsesfejlene ud fra kendskabet til tidligere observerede forudsigelsesfejl. Metoden giver dog ikke stor forbedring ved længere forudsigelser.

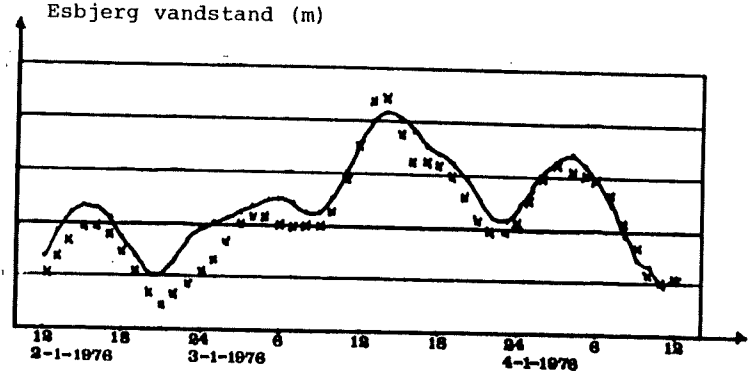
Variansen af forudsigelsesfejlene er ikke alene et udtømmende mål for, hvor god en forudsigelse er, men man bør eksempelvis også undersøge forudsigelsesfejlene ved den maksimale vandstand, som naturligvis er mest interessant i en given stormflodssituation. Der er derfor afbildet 1-times forudsigelser og de observerede vandstande nedenfor. På figuren p. 9.177 ses tilsvarende HN-modellens beregnede vandstande.



xxx : Observeret vandstand

——— : Forudsagt vandstand

1-timers forudsigelser med ARIMA-model.



XXX: Observeret vandstand

——: Forudsagt vandstand

HN-modellens beregnede vandstande.

Af disse og lignende figurer for 2-6 timers forudsigelserne fremgår, at den her anvendte metode er velegnet til forbedring af HN-modellens korttidsforudsigelser.

Ved vurdering af disse resultater må det imidlertid tages i betragtning, at undersøgelsen er af foreløbig karakter, men resultaterne indicerer, at man med de foreslåede modeller vil kunne opnå yderligere forbedringer bl.a. ved en mere minutiøs analyse af "sæson"-problemet (og dette kan delvis ske ved en nøjere analyse af ARIMA-processens residualer).

□

Referencer til kapitel 9

Alavi, A.S., & G.M. Jenkins: An example of digital filtering. Applied Statistics, vol. 14, 1965, p. 70.

Anderson, T.W.: The Statistical Analysis of Time Series. John Wiley & Sons, New York 1971.

Box, G.E.P., & Jenkins, G.: Time Series Analysis. Forecasting and Control. Holden-Day, San Francisco 1970.

Brown, R.G.: Smoothing, Forecasting and Prediction of Discrete Time Series. Prentice Hall, New Jersey 1963.

Chakravarti, I.M., R.G. Laha & J. Roy: Handbook of Methods of Applied Statistics. Vol. I. John Wiley & Sons, New York 1967.

Clelland, R.C., J.S. deCani & F.E. Brown: Basic Statistics with Business Applications. John Wiley & Sons, New York 1973.

Conradsen, K., N.H. Hansen, P.M. Larsen & H. Spliid: Forbedring af modeller til forudsigelse af extreme vandstande i Vadehavet. IMSOR 1977.

Dahlgaard, P.: Statistical Aspects of Tide Prediction. Licentiatafhandling, IMSOR 1973.

Gregg, J.V., C.H. Hossell & J.T. Richardson: Mathematical Trend Curves: An Aid to Forecasting. Oliver & Boyd, Edinburgh 1964.

Grenander, V., & M. Rosenblatt: Statistical Analysis of Stationary Time Series. John Wiley & Sons, New York 1957.

Hannan, E.J.: Multiple Time Series. John Wiley & Sons, New York 1970.

Hansen, N.H.: Problemer ved forudsigelser af lydhastighed i de danske farvande. Analyse af et stokastisk system. Licentiatafhandling, IMSOR 1967.

Jenkins, G.M., & D.G. Watts: Spectral analysis and its applications. Holden-Day, San Francisco 1968.

Kendall, M.: Time-Series (2nd. ed.). Charles Griffin and Co. Ltd, London 1976.

Kendall, M., & A. Stuart: The Advanced Theory of Statistics, vol. III. Charles Griffin and Co. Ltd., London 1966.

Lighthill, M.J.: Introduction to Fourier Analysis and Generalized Functions. Cambridge University Press, Cambridge 1958.

Lyck, E., & Gryning, S.-E.: Undersøgelse og behandling af svovldioxid- og svævestøvsmålinger foretaget i København 1968-71. Eksamensprojekt, Laboratoriet for Varme- og Klimateknik, DtH, 1972.

Nelson, C.R.: Applied Time Series Analysis For Managerial Forecasting. Holden-Day, San Francisco 1973.

Papoulis, A.: The Fourier Integral and Its Applications. McGraw-Hill, New York 1962.

Spliid, H.: En statistisk model for stormflodsvarsling. Licentiatafhandling, IMSOR 1973.

Tukey, J.W.: An Introduction to the Calculations of Numerical Spectrum Analysis. I Harris, B. (ed.): Spectral Analysis of Time Series. John Wiley & Sons, New York 1967.

Wold, H.: Bibliography on Time Series and Stochastic Processes. Oliver and Boyd, Edinburgh 1965.

INDEX

a posteriori fordeling, 7.2
a priori fordeling, 7.2
affin afbildning, 1.9
affin støtte, 2.13
algebra, 1.1 f
aliasing, 9.27
alignment, 9.143
Andersons U, 6.19
ANOVA, 5.69, 5.78 f
AR-proces,
  kontinuert, 9.61 f
  diskret, 9.39, 9.51 f, 9.92, 9.97
ARMA-proces
  kontinuert, 9.61
  diskret, 9.57 f
ARIMA-proces, 9.152 f
  multiplikativ sæsonmodel, 9.156 f
associativ lov, 1.2, 1.3
autokorrelationsfunktion, 9.31
  AR-proces, diskret, 9.52
  MA-proces, 9.49, 9.165
Autokovariansfunktion, 9.31 f
  ARMA-proces, 9.58
  estimation, 9.63 f
  hvid støj, 9.44, 9.59
  lineær proces, diskret, 9.47
  lineær proces, kontinuert, 9.60
  stationær proces, 9.32
autoregressiv proces, se AR-proces

B, 9.15
backwards elimination, 4.32 f, 4.42
bagudrettet forskydningsoperator, 9.15
balanceret variansanalyse, 5.13
Bartlett vindue, 9.5, 9.108, 9.113, 9.118
basis, 1.5
Bayes løsning, se diskriminantanalyse
beregningsformler, 3.58
beslutningsfunktion, 7.2
betinget fordeling, 2.22 f, 2.49 f
betinget middelværdi givet proces, 9.160
bibetingelser, 3.25
blokmatrix, 1.19 f
Box-Jenkins' metode, 9.152 f
båndpas filter, 9.117

"canonical variables, the first two", 7.34
central grænseværdisætning, 2.24
Cholesky faktorisering, 1.39
cofactor, 1.18
cospektrum, 9.138, 9.141
Cramér-Rao's ulighed, 2.65
Cramér's sætning, 1.20
cyklisk svingning, 9.81 f

δ, se Dirac's δ
datavindue, 9.24
definit, 1.46
determinant, 1.17 f, 1.20, 1.21, 1.40
deterministisk proces, 9.43
diag, 1.11
diagonalelement, 1.11
differensligningsform, 9.154
differensoperator, 9.15
differentiation af
  kvadratisk form, 1.60
  linearform, 1.60
digital filtrering af tidsrække, 9.120 f
dim, 1.5
dimension, 1.5
Dirac's δ, 9.9 f
direkte sum, 1.6 f, 1.68
discriminant score, 7.25
diskriminantanalyse, 7.1 f
  Bayes løsning, 7.1 f, 7.24 f, 7.26 f

diskriminantanalyse fortsat
flere normale populationer, 7.26 f, 7.32 f
flere populationer, 7.24 f
minimax løsning, 7.1 f
to normale populationer, 7.4 f
to populationer, 7.1 f
ukendte parametre, med, 7.15 f, 7.31, 7.32 f
diskriminantfunktion, diskriminator, 7.7, 7.32
diskriminantværdi, 7.25
dispersionsmatrix, 7.3 f
dispersionsmatrix, estimation af, se estimation
dispersionsmatrix, test for, se test
distributiv lov, 1.3
drejning, 1.40

Eckart-Youngs sætning, 1.42
effekt, 5.1
egenvektor, 1.36 f, 1.45
egenvektor m.h.t. matrix, 1.54
egenværdi, 1.35 f, 1.45, 1.58, 1.59
egenværdiproblem, generelle, 1.54 f
egenværdi m.h.t. matrix, 1.54
ellipsoide, 1.49
elliptisk cylinder, 1.51
empirisk dispersionsmatrix, 2.26
empirisk generaliseret varians, 2.63
empirisk partiel korrelation, 2.41
enhedsmatrix, 1.11
enhedsoperator, 9.16
ensemble, 9.30
equimax, 8.55
ergodeproblem, 9.64
estimation af/i
AR-proces, 9.166
ARIMA-proces, 9.164 f
dispersionsmatrix, 2.25, 2.61, 6.2, 6.8, 6.16 f, 7.15
egenværdi i dispersionsmatrix, 8.7
faktor vægte, 8.24 f
faktor værdi, 8.33 f
flerdimensional generel lineær model, 6.15 f
flerdimensionale parametre, 2.64 f
flerdimensional variansanalyse, 6.29, 6.33
generel lineær model, 3.5 f
kanoniske korrelationer, 8.20
kanoniske variable, 8.20
kovariansfunktion, 9.63 f
krydskovariansfunktion, 9.136 f
krydsspektrum, 9.141 f
MA-proces, 9.166
multipel korrelationskoefficient, 2.47
normal fordeling, 2.25
partiel korrelationskoefficient, 2.37
principal komponent, 8.7
spektrum, 9.100 f
euklidisk afstand, 1.65
exponentiel udjævning, 9.73 f, 9.125 f

F, 9.15
factor loading, 8.21
factor score, 8.21
faktor, 5.1
faktor vægt, 8.21
faktor værdi, 8.21
faktor analyse, 8.20 f
estimation af vægte, 8.24 f
maximum likelihood analyse, 8.48 f
principal faktorløsning, 8.27
rotation, 8.27 f
Q-modus analyse, 8.51 f
test for model, 8.50 f
faktorer, fælles, 8.21
faktorer, unikke, 8.22
fasefunktion, 9.20
fasespektrum, 9.138
filtrering, 9.116 f
flerdimensional generel lineær model, se generel lineær model, flerdimensional
flerdimensional normal fordeling, se normal fordeling, flerdimensional
flerdimensional variansanalyse, se variansanalyse, flerdimensional
foldning, 9.27

foldningsintegral, 9.14
forskydningsoperator, 9.14 f
forstærkning, 9.20
forudsigelse, 9.36, 9.38, 9.41, 9.43
  i ARIMA-proces, 9.158 f
  opdatering af, 9.163
  v.h.a. exponentiel udjævning, 9.75
forudsigelsesligning, 9.160
forventningsværdi, 2.2 f
forward selection, 4.34 f
Fourier transformation, diskret, 9.3 f, 9.29
  kontinuert, 9.2 f
frekvensresponsfunktion, 9.20, 9.121
funktionel relation, 4.52
fysisk realiserbarhed, 9.24

gain function, 9.20
Gauss-Markov's sætning, 3.7, 6.16
generaliseret varians, 2.59, 2.63
generel lineær model, 3.1 f
generel lineær model, flerdimensional, 6.13 f
geodæsi, 3.32
glidende gennemsnits proces, se MA-proces
glidende gennemsnits udjævning, 9.68 f, 9.124 f
Gompertz' trend, 9.79

Hamming vindue, 9.7, 9.108, 9.113, 9.118
Hamming tapering, 9.122, 9.146
Hanning vindue, 9.6, 9.108, 9.113, 9.118, 9.143
Hanning tapering, 9.122, 9.125, 9.129
Heavyside funktion, 9.10
hierarkisk klassifikation, se variansanalyse, hierarkisk klassifikation
Hotellings T^2,
  enstikprøvesituation, 6.1 f
  tostikprøvesituation, 6.8 f, 6.43, 7.16
hvid støj, 9.44, 9.159, 9.91
høj-pas filter, 9.117, 9.123, 9.124, 9.125, 9.126

$\underset{\equiv}{I}$, 1.11
$\underset{\equiv}{I}_n$, 1.11
idempotent afbildning, 1.8
idempotent matrix, 1.13, 1.59
impulsmoduleret signal, 9.26
impulsresponsfunktion, 9.20, 9.118
indeterministisk proces, 9.43
indre produkt, 1.64 f
informationsmatrix, 2.65
invers form, 9.155
invers matrix, 1.13, 1.19, 1.21, 1.39, 1.64
inverst element, 1.2
invertibel proces, 9.47, 9.49, 9.51, 9.57
isomorfi, 1.10

kanonisk korrelationskoefficient, 8.17 f
kanonisk variabel, 8.17 f
karakteristisk ligning, 9.23, 9.49, 9.52, 9.62
klassifikation, se diskrimination
knude, 4.70
koherensspektrum, 9.139, 9.140
kommunalitet, 8.23
kommutativ lov, 1.2, 9.16
komplement, 1.18
konfidensinterval for
  forudsagt værdi, 3.32 f, 9.163
  korrelationskoefficient, 2.43
  partiel korrelationskoefficient, 2.43
konfidensområde for
  middelværdi, 6.6, 6.10
  spektrum, 9.113
konjugerede retninger, 1.67
konjugerede vektorer, 1.55
konturellipsoide, 2.18 f, 2.21
koordinater, 1.6
koordinattransformation, 1.13 f
koordinattransformationsmatrix, 1.14
korrelationskoefficient, 2.27, 2.30
korrelationsmatrix, 2.3
korrespondanceanalyse, 8.54
kovarians, 2.6 f
kovariansmatrix, 2.3 f

Kroneckerprodukt, 1.63
krydsamplitudespektrum, 9.138
krydsklassifikation, se variansanalyse, krydsklassifikation
krydskorrelationsfunktion, 9.134 f
krydskovariansfunktion, 9.133 f
  estimation, 9.136
krydsspektralanalyse, 9.132 f
krydsspektrum, 9.137 f
  estimation, 9.141 f
kvadratisk form, 1.46, 1.60
kvadratisk matrix, 1.11
kvadraturspektrum, 9.138
kvotienttest, 3.42

$\mathfrak{E}$, 9.15
Λ, se Wilk's Λ
lag, 9.15
lagvindue, 9.105
landmåling, 3.32
lav-pas filter, 9.117, 9.123, 9.124, 9.125, 9.126
linearkombination, 1.4
lineær afbildning, 1.8 f, 1.12 f
lineær afhængighed, 1.5
lineær funktionel relation, 4.52
lineær ligning, løsning af, 1.19, 1.27
lineær proces, 9.44 f, 9.59 f
lineær regressionsanalyse, 4.1 f
lineær uafhængighed, 1.4
lineært bånd, 3.13, 3.25
lineært system, tidsinvariant, 9.19 f, 9.44, 9.60, 9.94, 9.135, 9.140
Little Jiffy, 8.50
logaritmisk parabolsk trend, 9.79
logistisk kurve, 4.68
logistisk trend, 9.79
logit, 4.68

MA-proces, 9.48 f, 9.61, 9.91
Mahalanobis afstand, 7.16
Mahalanobis afstand, generaliseret, 7.37
matrix, 1.10 f
matrix, regulær, 1.13
matrixprodukt, 1.12
matrixsum, 1.11
maximum likelihood estimation, 2.68
MDISC, 7.36 f
middelkvadratafvigelsessum, forventet værdi af, 5.10, 5.13, 5.15, 5.25, 5.32, 5.43, 5.45, 5.48, 5.49, 5.51, 5.58 f, 5.60 f, 5.62
middelværdi, 2.1 f
middelværdifunktion, 9.30
modificeret exponentielt trend, 9.79
Moore-Penrose invers, 1.35
multicollinaritet, 4.57
multipel korrelationskoefficient, 2.44 f, 4.6

$N_p(\underline{\mu},\underline{\Sigma})$, 2.11
∇, 9.15
∇_s, 9.156
neutralt element, 1.2
norm, 1.65
normal fordeling, flerdimensional, 2.10 f
normal fordeling, todimensional, 2.27 f
normalligninger, 3.5
nulrum, 1.9
Nyqvist-frekvens, 9.28

oblimin rotation, 8.55
oblique rotation, 8.33
opdatering af forudsigelse, 9.163
ortogonal matrix, 1.37
ortogonal regression, 2.53, 4.50 f
ortogonal transformation, 1.40
ortogonale polynomier, 4.14 f, 9.76
ortogonale vektorer, 1.36, 1.65, 1.67
ortonormal basis, 1.37

parabolsk trend, 9.79
partiel autokorrelationsfunktion, 9.55 f
  AR-proces, 9.21, 9.165
  ARMA-proces, 9.59
  MA-proces, 9.56
partiel korrelationskoefficient, 2.35 f, 4.7

Parzen vindue, 9.8, 9.108, 9.113, 9.118, 9.143
positiv definit, 1.46 f
positiv semidefinit, 1.46 f
power spectrum, se spektrum
prediktion, 4.27, 4.56, 4.65, se også forudsigelse
prediktionsinterval, 3.32 f, 9.163
prewhitening, 9.67, 9.123
principale komponenter, 4.55, 8.2 f
principale koordinater, 8.54
prikprodukt, 1.64 f
"probability associated with largest discriminant function", 7.37
projektion, 1.8, 1.65
præcision, 2.13
pseudoinvers afbildning, 1.25 f
pseudoinvers matrix, 1.2, 1.22 f, 1.59, 1.64
pythagoræiske læresætning, 1.65

Q-modus, 1.42, 1.46, 8.51
quartimax rotation, 8.28

R^2, 4.6
R^n, 1.3, 1.5, 1.10
R-modus, 1.42, 1.46
random shock, 9.45
random shock form, 9.154
rang af afbildning, 1.15
rang af matrix, 1.16, 1.43, 1.44
Rayleigh's kvotient, 1.49
regression, 2.49 f
regressionsanalyse, 4.1 f
  efter ortogonale polynomier, 4.14 f
  flerdimensional, 6.15, 6.23
  ikke lineær, 4.66 f

ortogonal regression, 4.50
ridge regression, 4.56 f

regressionsligning, valg af
  bedste, 4.27 f
  backwards elimination, 4.32 f, 4.42
  forward selection, 4.34 f, 4.41
  max R^2 improvement, 4.42
  samtlige regressioner, 4.30 f, 4.42
  stepwise regression, 4.38 f, 4.42
regulær matrix, 1.13, 1.16, 1.17
rektangulært vindue, 9.8, 9.108, 9.113, 9.118
reproduktivitetssætning for
  normal fordeling, 2.23
  Wishart fordeling, 2.61
residual, 3.10, 4.9, 9.81 f
residualplot, 4.9
retlinet trend, 9.79
rg, 1.16
ridge estimator, 4.59
ridge regression, 4.56 f
ridge trace, 4.62, 4.64
rækkevektor, 1.10

S, 9.16
S^2, forventet værdi, se middelkvadratafvigelsessum
SAK, 3.43, 6.30, 6.33
  beregning, 3.58, 5.52 f
  frihedsgrader, 5.56, 5.58
  forventet værdi, se middelkvadratafvigelsessum
sampling-problem, 9.24 f
semidefinit, 1.46
sideunderrum, 1.4
similaritetsmål, 8.54
similære matricer, 1.15
singulær værdi, 1.43
singulærværdi dekomposition, 1.4 f
sinusproces, 9.35 f, 9.96
  forstyrret, 9.38 f, 9.97
  skjult, 9.37 f, 9.97
skalainvarians, 8.48
skalamultiplikation, 1.2, 1.11
skalarprodukt, 1.64 f
Slutski's sætning, 9.128
spaltningssætning, 2.53 f
span, 1.4
spejling, 1.40
spektraldekomposition for matrix, 1.39
spektralfordeling, 9.90
spektralrepræsentation, 9.44
spektraltæthed, 9.88
spektralvindue, 9.24, 9.108, 9.113

spektrum, 9.88 f
AR-proces, 9.92, 9.97
estimation, 9.100 f
hvid støj, 9.91
integreret, 9.89
konfidensområde, 9.113
lineær proces, 9.95
MA-proces, 9.91
udglattet, 9.105
spektrum for matrix, 1.39
spline funktion, 4.70 f
spor af matrix, 1.58 f
stabilt system, 9.22
stationaritet, 9.32 f, 9.47, 9.49, 9.51, 9.57, 9.62
stepwise regression, 4.38 f
stikprøvespektrum, 9.100
stokastisk matrix, 2.1
stokastisk proces, 9.29 f
fordeling af, 9.30
støtte, 2.13
successiv testning, 3.48 f
summationsoperator, 9.16
symmetrisk matrix, 1.11, 1.37
sæsondifferensoperator, 9.156
sæsonindex, 9.82
sæsonvariation, 9.81 f, 9.156 f
søjlevektor, 1.10

T^2, se Hotelling's T^2
tabsfunktion, 7.2, 7.24
tendens, se trend
tensorprodukt, 1.63
test for/i
bedste diskriminantfunktion, 7.19
diagonalstruktur i dispersionsmatrix, 6.40, 8.17
egenværdi i dispersionsmatrix, 8.9
egenværdi i korrelationsmatrix, 8.9
ens dispersionsmatricer, 6.43
faktormodel, 8.50
flerdimensional general lineær model, 6.18 f
flerdimensional variansanalyse, 6.28 f
forudsætninger i regr. analyse, 4.8 f
general lineær model, 3.40 f
korrelationskoefficient, 2.42 f
middelværdi, 6.1 f, 6.8 f, 6.43, 7.16
multipel korrelationskoefficient, 2.48
partiel korrelationskoefficient, 2.42 f
proportional dispersion, 6.41
uafhængighed, 6.40, 8.17
yderligere information i diskr.an., 7.22
tidsinvariant lineært system, se lineært system
tidsrække, 9.2
tidsrækkeanalyse, 9.1 f
tog af δ-funktioner, 9.12
tr, 1.59
transponering, 1.10, 1.12, 1.64
trend, 9.76 f, 9.81

U, 6.19
U(p,q,r), 6.19
uafhængighed, 2.18 f
udjævning, 9.67 f
ukorreleret, 2.8
underrum, 1.4

varians, biologisk, 5.39
variansanalyse, 5.1 f, 6.28 f
blandet model, 5.23 f
ensidet, 5.3 f
ensidet flerdimensional, 6.28 f, 7.34, 7.38
variansanalyse fortsat
flere faktorer, 5.52 f
hierarkisk klassifikation, 5.18 f, 5.24 f, 5.44 f, 5.56 f, 5.67
krydsklassifikation, 5.18 f, 5.31 f, 5.40 f, 5.52 f, 5.67
robusthed, 5.63 f
systematisk model, 5.3, 5.7 f, 5.22 f
tilfældig model, 5.3, 5.5, 5.11 f, 5.22 f
tosidet, 5.18 f
tosidet flerdimensional, 6.30 f
trefaktor, 5.39 f
type I, se systematisk

type II, se tilfældig
varianskomponent, se tilfældig
vekselvirkning, 5.31, 5.38, 5.53 f
variansfunktion, 9.31
variate difference method, 9.77
variation
inden for grupper, 5.15, 6.29, 7.32
mellem grupper, 5.15, 6.29, 7.32
mellem hovedgrupper, 5.25
mellem undergrupper i.f. hvd.gr., 5.25
spaltning af totale, 3.42, 3.57, 4.18, 5.15, 5.25, 5.31, 5.42, 5.43, 5.44, 5.48, 5.49, 5.51, 5.56, 6.29, 6.32
systematisk, 5.2
tilfældig, 5.3
variationsbredde, 4.12
varimax rotation, 8.28, 8.55
VC, 2.9
vektoraddition, 1.2
vektorrum, 1.2 f
vinkel, 1.67
vægtet regression, 4.2
vægtfunktion, 9.20, 9.121

$W(n, \underline{\Sigma})$, 2.60
Wilk's Λ, 6.19, 7.34
Wishart fordeling, 2.59 f
Wolds dekomposition, 9.44

Yule-Walker ligninger, 9.52