En Introduktion til Statistik

Bind 2

Knut Conradsen

6. udgave

Kgs. Lyngby 2002

IMM

Trykt på IMM

Forord

Dette er 6 udgave af noten til kurset Multivariat Statistik.

5 udgaven var den første med LaTeXlayout. I forhold til den er der i nærværende udgave rettet en del trykfejl.

Fejl og rettelser må stadig meget gerne indrapporteres.

Knut Conradsen og Bjarne Ersbøll (be@imm.dtu.dk)

Indhold

1 Resumé af den lineære algebra 1

1.1 Vektorrum . . . 2

1.1.1 Definition af vektorrum . . . 2

1.1.2 Direkte sum af vektorrum . . . 5

1.2 Lineære afbildninger og matricer . . . 5

1.2.1 Lineære afbildninger . . . 7

1.2.2 Matricer . . . 8

1.2.3 Lineære afbildningers matrixfremstillinger . . . 10

1.2.4 Koordinattransformation . . . 11

1.2.5 Rangen af en matrix . . . 13

1.2.6 Determinanten af en matrix . . . 14

1.2.7 Blokmatricer . . . 16

1.3 Pseudoinvers eller generaliseret invers matrix . . . 18

1.4 Egenværdiproblemer. Kvadratiske former . . . 28

1.4.1 Egenværdier og egenvektorer for symmetriske matricer . . . 29

1.4.2 Singulær-værdi dekomposition af vilkårlig matrix. Q- og R-modus-analyser . . . 34

1.4.3 Kvadratiske former og positivt definitte matricer . . . 37

1.4.4 Det generelle egenværdiproblem for symmetriske matricer . . . 43

1.4.5 Sporet af en matrix . . . 46

1.4.6 Differentiation af linearform og kvadratisk form . . . 47

1.5 Tensor- eller Kronecker produkt af matricer . . . 50

1.6 Indre produkter og normer . . . 51

2 Flerdimensionale variable 57

2.1 Momenter af flerdimensionale stokastiske variable . . . 57

2.1.1 Middelværdi . . . 57

2.1.2 Dispersionsmatricen . . . 58

2.1.3 Kovarians . . . 61

2.2 Den flerdimensionale normalfordeling . . . 64

2.2.1 Definition og simple egenskaber . . . 64

2.2.2 Uafhængighed og konturellipsoider . . . 69
2.2.3 Betingede fordelinger . . . 74
2.2.4 Reproduktivitetssætning og central grænseværdisætning . . . 74
2.2.5 Estimation af parametre i en flerdimensional normal fordeling 75
2.2.6 Den todimensionale normale fordeling . . . 77
2.3 Korrelation og regression . . . 83
2.3.1 Den partielle korrelationskoefficient . . . 83
2.3.2 Den multiple korrelationskoefficient . . . 90
2.3.3 Regression . . . 93
2.4 Spaltningssætningen . . . 96
2.5 Wishart fordelingen og den generaliserede varians . . . 102
2.6 Lidt om estimation af flerdimensionale parametre . . . 106

3 Den generelle lineære model **111**
3.1 Estimation i den generelle lineære model . . . 111
3.1.1 Modelformulering . . . 111
3.1.2 Estimation i det regulære tilfælde . . . 114
3.1.3 Tilfældet $\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x}$ singulær . . . 120
3.1.4 Estimation under bibetingelser . . . 130
3.1.5 Konfidensintervaller for forudsagte værdier. Prediktionsinterval . . . 135
3.2 Test i den generelle lineære model . . . 141
3.2.1 Test for lavere dimension af modelrum . . . 141
3.2.2 Successiv testning i den generelle lineære model . . . 147

4 Regressionsanalyse **157**
4.1 Lineær regressionsanalyse . . . 157
4.1.1 Notation og model . . . 157
4.1.2 Korrelation og regression . . . 161
4.1.3 Analyse af forudsætninger . . . 162
4.1.4 Noget om "Influence Statistics" . . . 165
4.2 Regression efter ortogonale polynomier . . . 170
4.2.1 Definition og modelformulering . . . 170
4.2.2 Bestemmelse af ortogonale polynomier . . . 174
4.3 Valg af den "bedste" regressionsligning . . . 180
4.3.1 Problemstillingen . . . 180
4.3.2 Undersøgelse af samtlige regressioner . . . 183
4.3.3 Backwards elimination . . . 184
4.3.4 Forward selection . . . 186
4.3.5 Stepwise regression . . . 189
4.3.6 Nogle eksisterende programmer . . . 191
4.3.7 Numerisk appendix . . . 192
4.4 Andre regressionsmodeller og -løsninger . . . 198

4.4.1 Ortogonal regression (lineær funktionel relation) . . . 198
4.4.2 Ridge-regression . . . 202
4.4.3 Ikke-lineær regression og kurvetilpasning . . . 207

5 Variansanalyser **215**
5.1 Indledning . . . 215
5.2 Ensidet variansanalyse . . . 216
5.2.1 Modeller . . . 217
5.2.2 Analyse af den systematiske model . . . 219
5.2.3 Analyse af den tilfældige model . . . 222
5.2.4 Resumé af analyserne og et eksempel . . . 225
5.3 Tosidet variansanalyse. Hierarkisk klassifikation og krydsklassifikation 227
5.3.1 Hierarkisk klassifikation og krydsklassifikation . . . 227
5.3.2 Analyse af hierarkisk klassificerede data . . . 232
5.3.3 Analyse af krydsklassificerede data . . . 237
5.4 Variansanalysemodeller med 3 faktorer . . . 242
5.5 Variansanalyser med flere faktorer . . . 251
5.5.1 Estimation af parametre og beregning af kvadratafvigelsessummer . . . 252
5.5.2 Beregning af forventede værdier af middel-kvadratafvigelsessummer . . . 256

6 Test i den flerdimensionale normale fordeling **273**
6.1 Test for middelværdier . . . 273
6.1.1 Hotelling's T^2 i enstikprøvesituationen . . . 273
6.1.2 Hotelling's T^2 i tostikprøvesituationen . . . 279
6.2 Den flerdimensionale generelle lineære model . . . 283
6.3 Variansanalyser for flerdimensionale variable . . . 294
6.3.1 Ensidet flerdimensional variansanalyse . . . 294
6.3.2 Tosidet flerdimensional variansanalyse . . . 296
6.4 Tests vedrørende dispersionsmatricer . . . 303
6.4.1 Tests vedrørende en enkelt dispersionsmatrix . . . 303
6.4.2 Test for, at flere dispersionsmatricer er ens . . . 306

7 Diskriminantanalyse **309**
7.1 Diskrimination mellem 2 populationer . . . 309
7.1.1 Bayes- og minimaxløsninger . . . 309
7.1.2 Diskrimination mellem 2 normale populationer . . . 312
7.1.3 Diskrimination med ukendte parametre . . . 321
7.1.4 Test for bedste diskriminantfunktion . . . 323
7.1.5 Test for yderligere information . . . 326
7.2 Diskrimination mellem flere populationer . . . 327
7.2.1 Bayesløsning . . . 327
7.2.2 Bayesløsning i tilfældet med flere normale fordelinger . . . 329

7.2.3 Alternativ diskriminationsprocedure i tilfældet flere populationer 334
7.3 Nogle standardprogrammer til beregning af lineære diskriminatorer . 338

8 Principale komponenter kanoniske variable og korrelationer og faktoranalyse 343
8.1 Principale komponenter 344
8.1.1 Definition og simple egenskaber 344
8.1.2 Estimation og testning 349
8.2 Kanoniske variable og kanoniske korrelationer 355
8.3 Faktoranalyse 358
8.3.1 Model og forudsætninger 358
8.3.2 Estimation af faktorer (faktorvægte) 361
8.3.3 Faktor rotation 364
8.3.4 Beregning af faktorværdier (factor scores) 368
8.3.5 Et case-study 372
8.3.6 Lidt om maximum likelihood faktoranalyse 381
8.3.7 Q-modus analyse 383
8.3.8 Nogle standardprogrammer 386

A Det græske alfabet 397

Kapitel 1

Resumé af den lineære algebra

I dette kapitel skal vi give en oversigt over den lineære algebra med særlig henblik på dens anvendelse i statistikken. Kapitlet er ikke tænkt som en indføring i emnet, men kun som en genopfriskning af kendt stof. Der vil således ikke blive givet særligt mange eksempler indenfor de områder, der sædvanligt dækkes af kurser i algebra og geometri. Indenfor emner, der sædvanligvis ikke underkastes en så nøje gennemgang ved all round kurser, men som skønnes at have væsentlig betydning for en succesfuld anvendelse af algebraen i statistikken, vil der selvsagt blive givet flere eksempler og af og til også beviser.

I de senere år er man begyndt at involvere begreb som duale vektorrum i teorien for flerdimensional normal analyse. Uagtet de fordele dette medfører, har forfatteren dog valgt ikke at følge denne linie, hvorfor emnet selvsagt heller ikke berøres i dette oversigtskapitel.

I forbindelse med analyse af flerdimensionale statistiske problemer kommer man ofte ud for at skulle invertere matricer, der ikke nødvendigvis er regulære. Dette vil f.eks. være tilfældet, hvis man betragter et problem, der er givet på et ægte underrum i det betragtede n-dimensionale vektorrum. I stedet for at reducere betragtningerne til det relevante underrum foretrækker mange (= de fleste) forfattere at give delvist algebraiske løsninger ved at indføre den såkaldte pseudo-inverse til en ikke-regulær matrix. For at lette læsningen af (tidsskrift-) litteraturen vil dette begreb derfor blive omtalt og søgt anskueliggjort geometrisk.

Her må det også præciseres, at anvendelsen af pseudo inverse matricer giver en ganske bekvem måde at opstille løsninger til diverse matrixligninger i en algoritmisk form.

1.1 Vektorrum

Vi indleder med at give en oversigt over definitionen af og elementære egenskaber ved det helt fundamentale begreb, et (lineært) vektorrum.

1.1.1 Definition af vektorrum

Et **vektorrum (over de reelle tal)** er en mængde V forsynet med en kompositionsregel $+$ i mængden $V \times V \to V$ kaldet **vektoraddition** og en kompositionsregel $\cdot$ i $R \times V \to V$ kaldet **skalarmultiplikation**, som opfylder

i) $\forall \mathbf{u}, \mathbf{v} \in V: \quad \mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$ (kommutative lov for vektoraddition)

ii) $\forall \mathbf{u}, \mathbf{v}, \mathbf{x} \in V: \quad \mathbf{u} + (\mathbf{v} + \mathbf{x}) = (\mathbf{u} + \mathbf{v}) + \mathbf{x}$ (associative lov for vektoraddition)

iii) $\exists \mathbf{0} \in V \forall \mathbf{u} \in V: \quad \mathbf{u} + \mathbf{0} = \mathbf{u}$ (eksistens af neutralt element)

iv) $\forall \mathbf{u} \in V \exists - \mathbf{u} \in V: \mathbf{u} + (-\mathbf{u}) = \mathbf{0}$ (eksistens af invers element)

v) $\forall \lambda \in R \forall \mathbf{u}, \mathbf{v} \in V: \quad \lambda(\mathbf{u} + \mathbf{v}) = \lambda \mathbf{u} + \lambda \mathbf{v}$ (distributive lov for skalar multiplikation)

vi) $\forall \lambda_1, \lambda_2 \in R \forall \mathbf{u} \in V: \quad (\lambda_1 + \lambda_2)\mathbf{u} = \lambda_1 \mathbf{u} + \lambda_2 \mathbf{u}$ (distributive lov for skalarmultiplikation)

vii) $\forall \lambda_1, \lambda_2 \in R \forall \mathbf{u} \in V: \quad (\lambda_1 \lambda_2)\mathbf{u} = \lambda_1(\lambda_2 \mathbf{u})$ (associative lov for skalarmultiplikation)

viii) $\forall \mathbf{u} \in V: \quad 1\mathbf{u} = \mathbf{u}$

EKSEMPEL 1.1. Det eftervises let, at alle ordnede n-tupler

$$\mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$$

af reelle tal udgør et vektorrum, hvis kompositionerne defineres elementvis, i.e.

$$\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \end{bmatrix}$$

og

$$\lambda \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} \lambda x_1 \\ \vdots \\ \lambda x_n \end{bmatrix}$$

Dette vektorrum betegnes R^n ♦

Et vektorrum U som er en delmængde af et vektorrum V kaldes et **underrum** i V. Betragter vi omvendt vektorer $\mathbf{v}_1, \ldots, \mathbf{v}_k \in V$, kan vi definere

$$\text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$$

som det mindste underrum af V, der indeholder $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$. Det vises let, at

$$\text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\} = \{\sum_{i=1}^{k} \alpha_i \mathbf{v}_i | \alpha_i \in R, \quad i = 1, \ldots, k\}.$$

En vektor af formen $\sum \alpha_i \mathbf{v}_i$ kaldes en **linearkombination** af vektorerne $\mathbf{v}_i$, $i = 1, \ldots, k$. Ovenstående resultat kan da udtrykkes, at $\text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ netop består af alle linearkombinationer af vektorerne $\mathbf{v}_1, \ldots, \mathbf{v}_k$. Generelt defineres

$$\text{span}(U_1, \ldots, U_p)$$

hvor $U_i \subseteq V$, som det mindste underrum af V, der indeholder alle U_i, $i = 1, \ldots, p$.

Ved et **sideunderrum** forstås en mængde af formen

$$\mathbf{v} + U = \{\mathbf{v} + \mathbf{u} | \mathbf{u} \in U\},$$

hvor U er et underrum i V.

Situationen er skitseret i fig. 1.1.

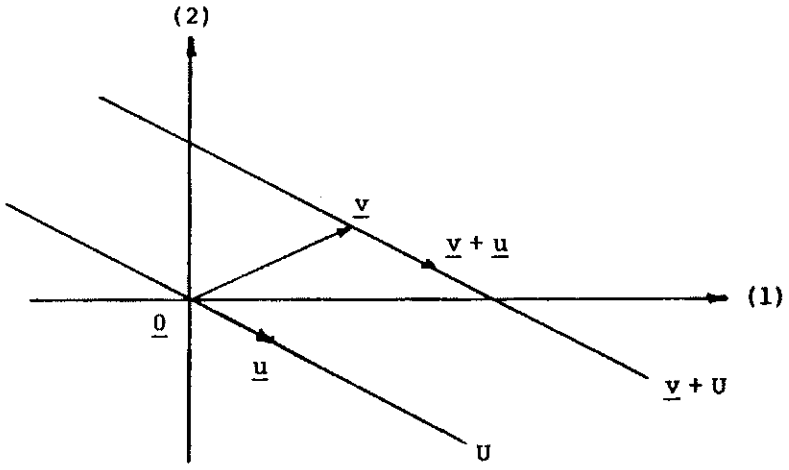
Vektorer $\mathbf{v}_1, \ldots, \mathbf{v}_n$ siges at være **lineært uafhængige**, hvis relationen

$$\alpha_1 \mathbf{v}_1 + \cdots + \alpha_n \mathbf{v}_n = 0$$

medfører, at

$$\alpha_1 = \cdots = \alpha_n = 0$$

I modsat fald siges de at være **lineært afhængige**, og mindst en af dem kan da skrives som en linearkombination af de øvrige.
En **basis** for vektorrummet V er en mængde af lineært uafhængige vektorer, som udspænder hele V. En vilkårlig vektor har en entydig fremstilling som linearkombination af vektorer i en basis. Antallet af elementer i forskellige baser for et vektorrum er altid det samme. Hvis dette antal er endeligt kaldes det vektorrummets **dimension** og det skrives $\dim(V)$.



Figur 1.1: Underrum og tilhørende sideunderrum i R^2.

EKSEMPEL 1.2. R^n har basen

$$\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \ldots, \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix},$$

og er således n-dimensional. ♦

I en fremstilling

$$\mathbf{v} = \sum_{i=1}^{n} \alpha_i \mathbf{v}_i$$

hvor $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ er en basis for V, kaldes sættet $\alpha_1, \ldots, \alpha_n$ $\mathbf{v}$'s **koordinater** med hensyn til basen $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$.

1.1.2 Direkte sum af vektorrum

Lad V være et (endelig dimensionalt) vektorrum og lad $U_1, \ldots, U_k$ være underrum af V. Vi siger da, at V er den **direkte sum af underrummene** $U_1, \ldots, U_k$, og vi skriver

$$V = U_1 \oplus \cdots \oplus U_k = \bigoplus_{i=1}^{k} U_i,$$

hvis en vilkårlig vektor $\mathbf{v} \in V$ har netop én fremstilling af formen

$$\mathbf{v} = \mathbf{u}_1 + \cdots + \mathbf{u}_k, \quad \mathbf{u}_1 \in U_1, \ldots, \mathbf{u}_k \in U_k \tag{1.1}$$

Denne betingelse er ensbetydende med, at der for vektorer $\mathbf{u}_i \in U_i$ gælder

$$\mathbf{u}_1 + \cdots + \mathbf{u}_k = \mathbf{0} \quad \Rightarrow \quad \mathbf{u}_1 = \cdots = \mathbf{u}_k = 0$$

Dette er igen ensbetydende med, at

$$\dim(\operatorname{span}(U_1, \ldots, U_k)) = \sum_{i=1}^{k} \dim U_i = \dim V$$

Dette er endelig ensbetydende med, at alle fællesmængder af nogle af U_i'erne er $\mathbf{0}$. Det er selvfølgelig et gennemgående krav, at $\operatorname{span}(U_1, \ldots, U_k) = V$, d.v.s. at der overhovedet findes en fremstilling som 1.1. Det er 1.1's eventuelle entydighed, der bevirker, at vi eventuelt kalder "summen" direkte.

Vi skitserer nogle eksempler i nedenstående fig. 1.1.2

Er V spaltet i en direkte sum

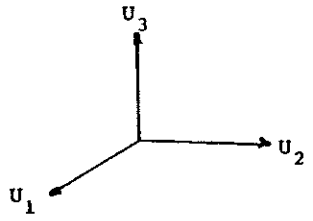
$$V = U_1 \oplus \cdots \oplus U_k$$

kalder vi en vilkårlig vektor $\mathbf{v}$'s komponent i U_i for $\mathbf{v}$'s **projektion** på U_i (efter retningen bestemt ved $U_1, \ldots, U_{i-1}, U_{i+1}, \ldots, U_k$) og vi benævner den $p_i(\mathbf{v})$

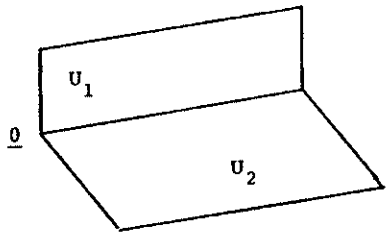
Afbildningen p_i er **idempotent**, i.e. $p_i \circ p_i(\mathbf{v}) = p_i(\mathbf{v}), \forall \mathbf{v}$ hvor f $\circ$ g betegner sammensætningen af f og g.

1.2 Lineære afbildninger og matricer

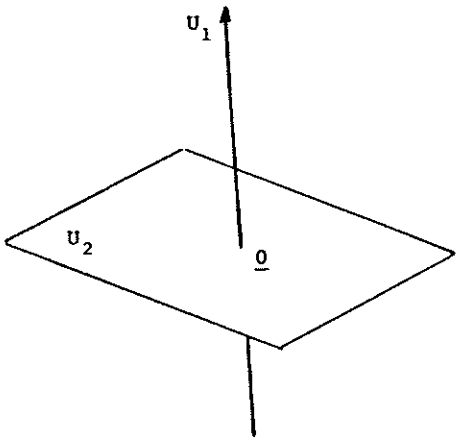
Vi indleder med et afsnit om lineære afbildninger



$U_1 \oplus U_2 \oplus U_3 = R^3$ Summen er direkte, e.g. fordi $\dim U_1 + \dim U_2 + \dim U_3 = 3$

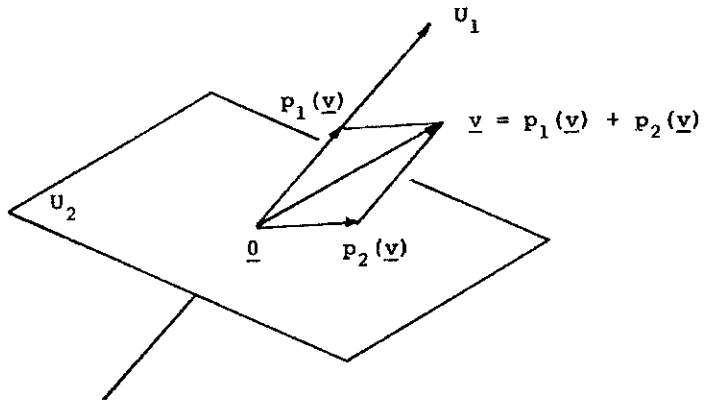


R^3 er ej direkte sum af U_1 og U_2; thi $\dim U_1 + \dim U_2 = 4$



Her er $U_1 \oplus U_2 = R^3$ e.g. fordi U_1 og U_2 foruden at udspænde R^3 også tilfredsstiller $U_1 \cap U_2 = \mathbf{0}$

Figur 1.2:



Figur 1.3: Projektion af vektor.

1.2.1 Lineære afbildninger

En afbildning $A : U \to V$, hvor U og V er vektorrum, siges at være lineær, hvis

$$\forall \lambda_1, \lambda_2 \in R \forall \mathbf{u}_1, \mathbf{u}_2 \in U : A(\lambda_1 \mathbf{u}_1 + \lambda_2 \mathbf{u}_2) = \lambda_1 A(\mathbf{u}_1) + \lambda_2 A(\mathbf{u}_2)$$

EKSEMPEL 1.3. En afbildning $A : R \to R$ er lineær, netop hvis dens graf er en ret linie gennem (0,0). Hvis grafen er en ret linie, der dog ikke går gennem (0,0) siges afbildningen at være **affin**.

Ved **nulrummet** $N(A)$ for en lineær afbildning $A : U \to V$ forstås underrummet

$$A^{-1}(\mathbf{0}) = \{\mathbf{u} | A(\mathbf{u}) = \mathbf{0}\}$$

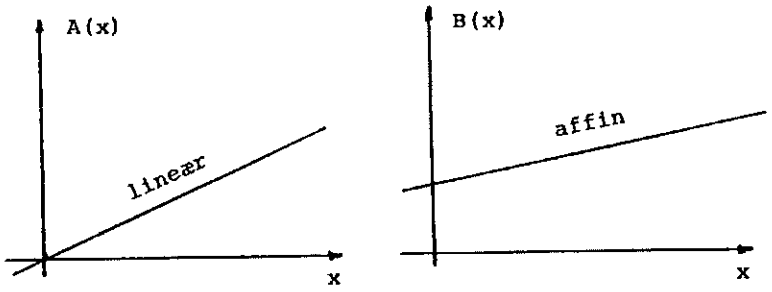
Der gælder følgende formel om sammenhængen mellem dimensionen af billedrum og nulrum

$$\dim N(A) + \dim A(U) = \dim U$$

Specielt gælder

$$\dim A(U) \leq \dim U$$

med lighedstegn netop, hvis A er injektiv (d.v.s. entydig). Er A bijektiv (d.v.s. entydig og "på"), ses umiddelbart, at $\dim U = \dim V$. Vi siger, at en sådan afbildning er en



Figur 1.4: Grafer for lineær og for affin afbildning $R \to R$.

isomorfi, og at U og V er **isomorfe**. Det kan vises, at **et vilkårligt n-dimensionalt (reelt) vektorrum er isomorft med R^n**. I den videre fremstilling vil vi derfor ofte identificere et n-dimensionalt vektorrum med R^n. ♦

Det kan vises, at de i forrige afsnit omtalte projektioner er lineære afbildninger.

1.2.2 Matricer

Ved en **matrix A** forstås et rektangulært skema af tal som

$$\mathbf{A} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mn} \end{bmatrix}.$$

Ofte vil vi anvende den kortere skrivemåde

$$\mathbf{A} = (a_{ij}).$$

Mere specifikt betegnes $\mathbf{A}$ som en $m \times n$ matrix, da der er m rækker og n søjler. Hvis $m = 1$ kan matricen kaldes en **rækkevektor** og hvis $n = 1$ en **søjlevektor**.

Den matrix man får frem ved at ombytte rækker og søjler kaldes den **transponerede** matrix til $\mathbf{A}$ og betegnes $\mathbf{A}'$, i.e.

$$\mathbf{A}' = \begin{bmatrix} a_{11} & \cdots & a_{m1} \\ \vdots & & \vdots \\ a_{1n} & \cdots & a_{mn} \end{bmatrix}$$

En $m \times n$ matrix kaldes **kvadratisk**, hvis $n = m$. En kvadratisk matrix for hvilken $\mathbf{A} = \mathbf{A}'$ kaldes **symmetrisk**. Elementerne a_{ii}, $i = 1, \ldots, n$ kaldes **diagonalelementerne**.

En særlig vigtig matrix er **enhedsmatricen** af orden n

$$\mathbf{I}_n = \mathbf{I} = \begin{bmatrix} 1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 1 \end{bmatrix}.$$

En matrix der har nuller uden for diagonalen kaldes en **diagonalmatrix**. Vi anvender skrivemåden

$$\mathbf{\Delta} = \operatorname{diag}(\delta_1, \ldots, \delta_n) = \begin{bmatrix} \delta_1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \delta_n \end{bmatrix}.$$

For givne $n \times m$ matricer $\mathbf{A}$ og $\mathbf{B}$ definerer man **matrix-summen**

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} a_{11} + b_{11} & \cdots & a_{1m} + b_{1m} \\ \vdots & & \vdots \\ a_{n1} + b_{n1} & \cdots & a_{nm} + b_{nm} \end{bmatrix}.$$

Skalarmultiplikation defineres ved

$$c\mathbf{A} = \begin{bmatrix} ca_{11} & \cdots & ca_{1m} \\ \vdots & & \vdots \\ ca_{n1} & \cdots & ca_{nm} \end{bmatrix},$$

d.v.s. elementvis multiplikation.

For en $m \times n$ matrix $\mathbf{C}$ og en $n \times p$ matrix $\mathbf{D}$ defineres **matrixproduktet** $\mathbf{P} = \mathbf{C}\,\mathbf{D}$ ved, at $\mathbf{P}$ er en $m \times p$ matrix med (i, j)'te element

$$p_{ij} = \sum_{k=1}^{n} c_{ik} d_{kj}$$

Det kan bemærkes, at matrixproduktet ikke er kommutativt, i.e. at $\mathbf{C}\,\mathbf{D}$ ikke er lig $\mathbf{D}\,\mathbf{C}$.

For transponeringen gælder nu følgende regler

$$\begin{aligned}(\mathbf{A}+\mathbf{B})' &= \mathbf{A}'+\mathbf{B}'\\ (c\mathbf{A})' &= c\mathbf{A}'\\ (\mathbf{C}\,\mathbf{D})' &= \mathbf{D}'\mathbf{C}'\end{aligned}$$

1.2.3 Lineære afbildningers matrixfremstillinger

Det kan vises, at der til enhver lineær afbildning $A : R^n \to R^m$ svarer en $m \times n$ **matrix A**, således at

$$\forall \mathbf{x} \in R^n : A(\mathbf{x}) = \mathbf{A}\,\mathbf{x}$$

Omvendt er et A defineret ved denne relation en lineær afbildning. $\mathbf{A}$ bestemmes let som den matrix, der som søjler har koordinaterne for billedet af enhedsvektorerne i R^n. E.g. haves

$$\mathbf{A}\,\mathbf{e}_2 = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} a_{12} \\ \vdots \\ a_{m2} \end{bmatrix} = \mathbf{a}_2$$

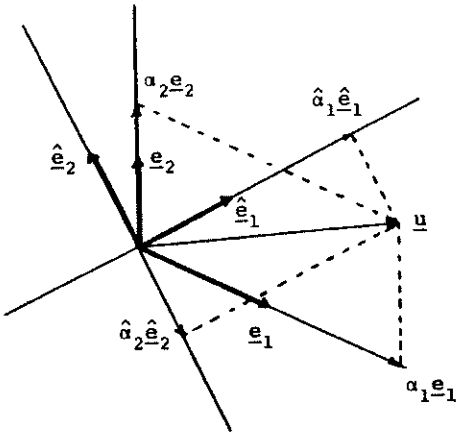
Er der ydermere givet en lineær afbildning $B : R^m \to R^k$ med tilhørende matrix $\mathbf{B}$ $(k \times m)$, da gælder, at $B \circ A \leftrightarrow \mathbf{B}\,\mathbf{A}$ d.v.s. :

$$\forall \mathbf{x} \in R^n (B \circ A(\mathbf{x}) = B(A(\mathbf{x})) = \mathbf{B}\,\mathbf{A}\,\mathbf{x})$$

Vi nævner her, at en $n \times n$ matrix $\mathbf{A}$ siges at være **regulær**, hvis den tilsvarende lineære afbildning er bijektiv. Dette er ensbetydende med eksistensen af en **invers matrix**, i.e. en matrix $\mathbf{A}^{-1}$, der tilfredsstiller

$$\mathbf{A}\,\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$$

hvor $\mathbf{I}$ er enhedsmatricen af n'te orden.



$$, \mathbf{u} = \begin{cases} \alpha_1\mathbf{e}_1 + \alpha_2\mathbf{e}_2 = \begin{bmatrix} \mathbf{e}_1 & \mathbf{e}_2 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} \\ \hat{\alpha}_1\hat{\mathbf{e}}_1 + \hat{\alpha}_2\hat{\mathbf{e}}_2 = \begin{bmatrix} \hat{\mathbf{e}}_1 & \hat{\mathbf{e}}_2 \end{bmatrix} \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{bmatrix} \end{cases}$$

Figur 1.5: Skitse over koordinattransformationsproblemet.

En kvadratisk matrix, der svarer til en idempotent afbildning, kaldes også **idempotent**. Det ses let, at en matrix $\mathbf{A}$ er idempotent, hvis og kun hvis

$$\mathbf{A}\,\mathbf{A} = \mathbf{A}$$

Det kan bemærkes, at hvis en idempotent matrix er regulær, da er den lig enhedsmatricen, d.v.s. den tilsvarende afbildning er identiteten.

1.2.4 Koordinattransformation

I dette afsnit vil vi angive formler for en lineær afbildnings matrixfremstilling ved overgang fra et sæt baser til et andet.

Vi betragter først ændringen i koordinater ved overgang fra et koordinatsystem til et andet. Normalt vælger vi at undlade at skelne mellem en vektor $\mathbf{u}$ og dens koordinatsæt. Det giver en enkel notation og vil ikke give anledning til forvekslinger. Når der er involveret flere koordinatsystemer, er det imidlertid åbenbart nødvendigt at foretage denne skelnen. Vi betragter i R^n to koordinatsystemer $(\mathbf{e}_1, \ldots, \mathbf{e}_n)$ og $(\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_n)$ Koordinaterne til en vektor $\mathbf{u}$ i de to systemer betegnes henholdsvis $(\alpha_1, \ldots, \alpha_n)'$ og $(\hat{\alpha}_1, \ldots, \hat{\alpha}_n)'$, jvf. figur 1.5.

Lad det "nye" system $(\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_n)$ være givet ved

$$(\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_n) = (\mathbf{e}_1, \ldots, \mathbf{e}_n)\mathbf{S}$$

d.v.s.

$$\hat{\mathbf{e}}_i = s_{1i}\mathbf{e}_1 + \cdots + s_{ni}\mathbf{e}_i, \qquad i = 1, \ldots, n.$$

Søjlerne i S-matricen er altså lig det "nye" systems "gamle" koordinater. S kaldes **koordinattransformationsmatricen.**

BEMÆRKNING 1.1. I mange fremstillinger anvendes udtrykket koordinattransformationsmatricen dog om matricen $\mathbf{S}^{-1}$. Man må derfor altid gøre sig klart, hvilken matrix der er tale om.

Da

$$(\mathbf{e}_1 \cdots \mathbf{e}_n) \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = (\hat{\mathbf{e}}_1 \cdots \hat{\mathbf{e}}_n) \begin{pmatrix} \hat{\alpha}_1 \\ \vdots \\ \hat{\alpha}_n \end{pmatrix},$$

(jvf. fig. 1.5), bliver sammenhængen mellem en vektors "gamle" koordinater og "nye" koordinater

$$\begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} = \mathbf{S} \begin{bmatrix} \hat{\alpha}_1 \\ \vdots \\ \hat{\alpha}_n \end{bmatrix} \Longleftrightarrow \begin{bmatrix} \hat{\alpha}_1 \\ \vdots \\ \hat{\alpha}_n \end{bmatrix} = \mathbf{S}^{-1} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}.$$

▼

Vi betragter nu en lineær afbildning $A : R^n \to R^m$, og lad A's matrixfremstilling m.h.t. baserne $(\mathbf{e}_1, \ldots, \mathbf{e}_n)$ og $(\mathbf{f}_1, \ldots, \mathbf{f}_m)$ være

$$\beta = \mathbf{A}\,\alpha$$

og fremstillingen m.h.t. baserne $(\hat{\mathbf{e}}_1, \ldots, \hat{\mathbf{e}}_n) = (\mathbf{e}_1, \ldots, \mathbf{e}_n)\mathbf{S}$ og $(\hat{\mathbf{f}}_1, \ldots, \hat{\mathbf{f}}_m) = (\mathbf{f}_1, \ldots, \mathbf{f}_m)\mathbf{T}$ være

$$\hat{\beta} = \hat{\mathbf{A}}\hat{\alpha}$$

Da gælder

$$\hat{\mathbf{A}} = \mathbf{S}^{-1}\mathbf{A}\,\mathbf{T},$$

hvilket fås umiddelbart ved anvendelse af koordinattransformationsreglerne for koordinaterne.

Er der tale om afbildninger $R^n \to R^n$ og anvendes samme koordinattransformation fås relationen

$$\hat{\mathbf{A}} = \mathbf{S}^{-1}\mathbf{A}\,\mathbf{S}.$$

Matricerne $\mathbf{A}$ og $\hat{\mathbf{A}} = \mathbf{S}^{-1}\mathbf{A}\,\mathbf{S}$ kaldes da **similære** matricer.

1.2.5 Rangen af en matrix

Ved **rangen** af en lineær afbildning $A : R^n \to R^m$ forstås dimensionen af billedrummet, i.e.

$$\mathrm{rg}(A) = \dim A(R^n).$$

Ved **rangen** af en **matrix** A forstås rangen af den tilsvarende lineære afbildning.

Det ses, at $\mathrm{rg}(\mathbf{A})$ netop er antallet af lineært uafhængige søjlevektorer i $\mathbf{A}$. Trivielt gælder derfor

$$\mathrm{rg}(\mathbf{A}) \leq n.$$

Indføres den transponerede matrix $\mathbf{A}'$ vises det let, at $\mathrm{rg}(\mathbf{A}) = \mathrm{rg}(\mathbf{A}')$, d.v.s. vi har

$$\mathrm{rg}(\mathbf{A}) \leq \min(m, n).$$

Hvis $\mathbf{A}$ og $\mathbf{B}$ er to $m \times n$ matricer, gælder

$$\mathrm{rg}(\mathbf{A} + \mathbf{B}) \leq \mathrm{rg}(\mathbf{A}) + \mathrm{rg}(\mathbf{B}).$$

Denne relation er umiddelbar, når man erindrer, at der for de tilhørende afbildninger A og B gælder $(A + B)\,(R^n) \subseteq A(R^n) \cup B(R^n)$.

Hvis $\mathbf{A}$ er en $(m \times n)$-matrix og $\mathbf{B}$ en $(k \times m)$-matrix, gælder

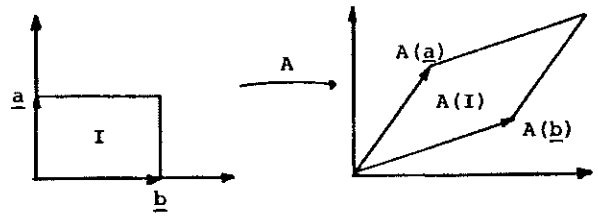
$$\mathrm{rg}(\mathbf{BA}) \leq \mathrm{rg}(\mathbf{A}).$$

Hvis $\mathbf{B}$ er **regulær** $(m \times m)$ gælder

$$\mathrm{rg}(\mathbf{BA}) = \mathrm{rg}(\mathbf{A}).$$

Disse relationer er umiddelbare følger af relationen $\dim B(A(R^n)) \leq \dim(A(R^n))$, hvor vi har lighedstegn, hvis B er injektiv. Der gælder naturligvis de analoge relationer for en $(n \times p)$-matrix $\mathbf{C}$:

$$\mathrm{rg}(\mathbf{A}\,\mathbf{C}) \leq \mathrm{rg}(\mathbf{A})$$



Figur 1.6: Et rektangel og dets billede ved en lineær afbildning.

med lighedstegn, hvis C er en regulær $(n \times n)$-matrix. Af disse udledes specielt for **regulære** B og C

$$\mathrm{rg}(\mathbf{B}\,\mathbf{A}\,\mathbf{C}) = \mathrm{rg}(\mathbf{A}).$$

Til slut nævner vi, at en $(n \times n)$-matrix A er regulær, netop hvis $\mathrm{rg}(\mathbf{A}) = n$.

1.2.6 Determinanten af en matrix

Den abstrakte definition på **determinanten** af en kvadratisk $p \times p$ matrix A er

$$\det(\mathbf{A}) = \sum_{\text{alle}\sigma} \pm a_{1\sigma(1)} \cdots a_{p\sigma(p)},$$

hvor σ er en permutation af tallene $1, \ldots, p$ og hvor + tegnet anvendes, hvis permutationen er lige (d.v.s. kan sammensættes af et lige antal ombytninger af naboer) og -, hvis den er ulige.

Vi skal ikke komme ind på baggrunden for denne definition, men uden bevis konstatere, at determinanten angiver volumenforholdet for den tilsvarende lineære afbildning, i.e. for en $(n \times n)$-matrix A

$$|\det(\mathbf{A})| = \frac{\mathrm{vol}(A(I))}{\mathrm{vol}(I)},$$

hvor I er en n-dimensional kasse og $A(I)$ billedet af I (et n-dimensionalt parallelepipedum) ved den tilsvarende afbildning.

I 2 dimensioner er situationen skitseret i fig. 1.6. For 2×2 og 3×3 matricer bliver definitionen på determinanten

$$\det\begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc$$

$$\det \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} = aei + bfg + cdh - gec - hfa - idb.$$

For determinanter af højere orden (her n'te) kan man udvikle determimanten efter den i'te række, d.v.s.

$$\det(\mathbf{A}) = \sum_{j=1}^{n} a_{ij}(-1)^{i+j} \det(\mathbf{A}_{ij}),$$

hvor $\mathbf{A}_{ij}$ er den matrix, der fremkommer ved at slette den i'te række og den j'te søjle i $\mathbf{A}$. Tallet

$$A_{ij} = (-1)^{i+j} \det(\mathbf{A}_{ij})$$

kaldes også for elementet a_{ij}'s **komplement** (eng. cofactor). Der eksisterer selvfølgelig en helt analog udvikling efter søjler.

Når man explicit skal evaluere determinanten er der tre "handy" regler:

i) ombytning af 2 rækker (søjler) i $\mathbf{A}$ multiplicerer $\det(\mathbf{A})$ med -1.

ii) multipliceres en række (søjle) med en skalar multipliceres $\det(A)$ med denne skalar.

iii) adderes et multiplum af en række (søjle) til en anden forbliver $\det(A)$ uændret.

Skal man bestemme rangen af en matrix, kan det være nyttigt at erindre, at rangen er det største tal r for hvilken matricen har en fra 0 forskellig underdeterminant af r'te orden. Vi får så specielt, at en matrix $\mathbf{A}$ er regulær, netop hvis det $\mathbf{A} \neq 0$. Dette virker også intuitivt indlysende, når man betragter determinanten som volumenforholdet. Er dette 0, må afbildningen være "dimensionsreducerende" i en passende forstand.

For kvadratiske matricer $\mathbf{A}$ og $\mathbf{B}$ gælder

$$\det(\mathbf{A}\,\mathbf{B}) = \det(\mathbf{A}) \det(\mathbf{B})$$

For en diagonalmatrix $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ gælder

$$\det(\mathbf{\Lambda}) = \lambda_1 \ldots \lambda_n$$

For en trekantmatrix $\mathbf{C}$ med diagonalelementer $c_1, \ldots, c_n$ gælder

$$\det(\mathbf{C}) = c_1 \cdots c_n$$

Ved hjælp af determinanter kan man direkte opskrive den inverse til en regulær matrix A. Der gælder nemlig, at

$$\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})}(A_{ij})',$$

d.v.s. den inverse til en regulær matrix $\mathbf{A}$ er den transponerede til den matrix, der fås ved at erstatte hvert element i $\mathbf{A}$ med sit komplement divideret med det $\mathbf{A}$. Det må dog her præciseres, at formlen ikke er direkte anvendelig ved inversion af store matricer grundet det store regnearbejde, der er forbundet med beregning af determinanter.

Noget tilsvarende gør sig gældende for **Cramérs** sætning om løsning af et lineært ligningssystem: Betragt den regulære matrix $\mathbf{A} = (\mathbf{A}_1, \ldots, \mathbf{A}_n)$. Da er løsningen til ligningen

$$\mathbf{A}\,\mathbf{x} = \mathbf{b}$$

givet ved

$$x_i = \frac{\det(\mathbf{a}_1, \ldots, \mathbf{a}_{i-1}, \mathbf{b}, \mathbf{a}_{i+1}, \ldots, \mathbf{a}_n)}{\det \mathbf{A}}$$

1.2.7 Blokmatricer

Ved **blokmatrix** forstås en matrix af formen

$$\mathbf{B} = \begin{bmatrix} \mathbf{B}_{11} & \cdots & \mathbf{B}_{1n} \\ \vdots & & \vdots \\ \mathbf{B}_{m1} & \cdots & \mathbf{B}_{mn} \end{bmatrix}$$

hvor **blokkene** $\mathbf{B}_{ij}$ er matricer af orden $m_i \times n_j$.

Ved addition og multiplikation kan man anvende de sædvanlige regneregler for matricer, og opfatte blokkene som elementer. Vi finder således eksempelvis, at

$$\begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{R} \\ \mathbf{S} \end{bmatrix} = \begin{bmatrix} \mathbf{A}\,\mathbf{R} + \mathbf{B}\,\mathbf{S} \\ \mathbf{C}\,\mathbf{R} + \mathbf{D}\,\mathbf{S} \end{bmatrix},$$

naturligvis forudsat, at de involverede produkter eksisterer etc.

Vi nævner nu først et resultat om determinanter af "trekants" matricen.

SÆTNING 1.1. Lad den kvadratiske matrix $\mathbf{A}$ være spaltet i blokmatricer

$$\mathbf{A} = \begin{bmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{O} & \mathbf{D} \end{bmatrix}$$

hvor $\mathbf{B}$ og $\mathbf{D}$ er kvadratiske og $\mathbf{O}$ en matrix bestående af lutter 0'er. Da gælder

$$\det(\mathbf{A}) = \det(\mathbf{B})\det(\mathbf{D})$$

▲

BEVIS 1.1. Vi har at

$$\begin{bmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{O} & \mathbf{D} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{O} \\ \mathbf{O} & \mathbf{D} \end{bmatrix} \begin{bmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{O} & \mathbf{I} \end{bmatrix}$$

hvor $\mathbf{I}$'erne er enhedsmatricer og ikke nødvendigvis af samme orden. Hvis man udvikler den første matrix efter 1.ste række, ses, at den har samme determinant som den matrix, man får ved at slette 1.ste række og søjle. Gentages dette indtil den tilovers-blevne underdeterminant er $\mathbf{D}$, ses, at

$$\det \begin{bmatrix} \mathbf{I} & \mathbf{O} \\ \mathbf{O} & \mathbf{D} \end{bmatrix} = \det(\mathbf{D})$$

Analogt fås, at den sidste matrix har determinanten det $\mathbf{O}$ og resultatet følger. ■

Den næste sætning udvider dette resultat.

SÆTNING 1.2. Lad matricen $\boldsymbol{\Sigma}$ være spaltet i blokmatricer

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

Da gælder

$$\det(\boldsymbol{\Sigma}) = \det(\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})\det(\boldsymbol{\Sigma}_{22}),$$

forudsat, at $\boldsymbol{\Sigma}_{22}$ er regulær. ▲

BEVIS 1.2. Da

$$\begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{O} \\ -\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} & \mathbf{I} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{12} \\ \mathbf{O} & \boldsymbol{\Sigma}_{22} \end{bmatrix},$$

følger resultatet umiddelbart ved anveldelse af foregående sætning. ■

Den sidste sætning giver et nyttigt resultat om inversion af matricer, der er spaltet i blokmatricer.

SÆTNING 1.3. For den symmetriske matrix

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

gælder

$$\Sigma^{-1} = \begin{bmatrix} \mathbf{B}^{-1} & -\mathbf{B}^{-1}\mathbf{A}' \\ -\mathbf{A}\mathbf{B}^{-1} & \Sigma_{22}^{-1} + \mathbf{A}\mathbf{B}^{-1}\mathbf{A}' \end{bmatrix},$$

hvor

$$\begin{aligned} \mathbf{A} &= \Sigma_{22}^{-1}\Sigma_{21}^{-1} \\ \mathbf{B} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}, \end{aligned}$$

hvor det naturligvis forudsættes, at de anførte inverse eksisterer. ▲

BEVIS 1.3. Resultatet følger umiddelbart ved multiplikation af Σ og Σ^{-1}. ■

1.3 Pseudoinvers eller generaliseret invers matrix til en ikke-regulær matrix

Vi betragter en lineær afbildning

$$A : E \to F$$

hvor E er et n-dimensionalt og F et m-dimensionalt (euklidisk) vektorrum. Den til A svarende matrix benævnes som sædvanlig $\mathbf{A}$ og den har dimensionen $m \times n$. Vi sætter nulrummet for A lig U, d.v.s.

$$U = A^{-1}(\mathbf{0}),$$

og benævner dets dimension r. Billedrummet

$$V = A(E)$$

har dimensionen $s = n - r$, jvf. p. 7.

Vi betragter nu et vilkårligt s-dimensionalt underrum $U^* \subseteq E$, som er komplementært til U, og et vilkårligt $m - s$ dimensionalt underrum $V^* \subseteq F$, som er komplementært til V.

En vilkårlig vektor $\mathbf{x} \in E$ kan nu skrives på formen

$$\mathbf{x} = \mathbf{u} + \mathbf{u}^*, \quad \mathbf{u} \in U \quad \text{og} \quad \mathbf{u}^* \in U^*,$$

idet $\mathbf{u}$ og $\mathbf{u}^*$ er givet ved

$$\begin{aligned} \mathbf{u} &= \mathbf{x} - p_{U^*}(x) \\ \mathbf{u}^* &= p_{U^*}(x) \end{aligned}$$

Her betegner p_{U^*} projektionen af E ind på U^* langs underrummet U. Tilsvarende kan ethvert $\mathbf{y} \in F$ skrives

$$\mathbf{y} = (\mathbf{y} - p_V(\mathbf{y})) + p_V(\mathbf{y}) = \mathbf{v}^* + \mathbf{v}$$

hvor

$$p_V : F \to V$$

er projektionen af F ind på V langs V^*

Da nu

$$A(\mathbf{x}) = A(\mathbf{u} + \mathbf{u}^*) = A(\mathbf{u}^*),$$

ses, at A er konstant på siderummene

$$\mathbf{u}^* + U = \{\mathbf{u}^* + \mathbf{u} | \mathbf{u} \in U\}$$

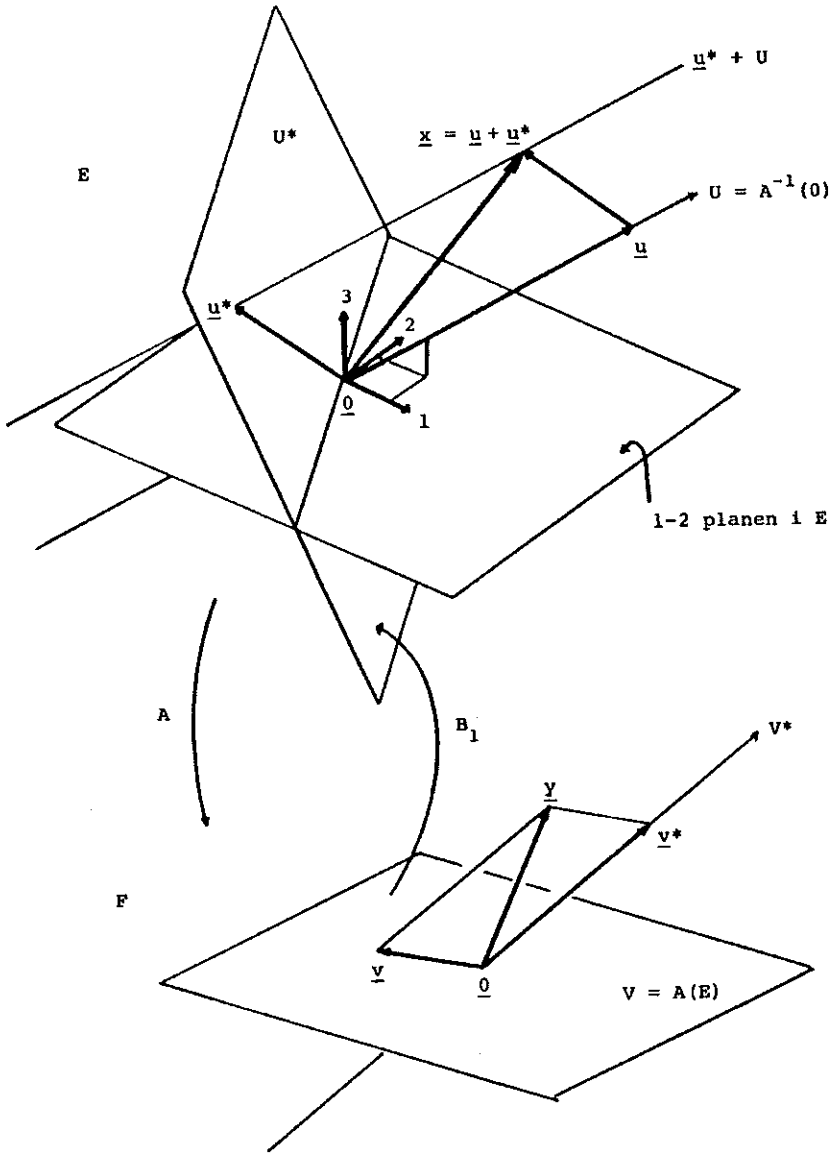
og det følger, at A's restriktion til U^* er en bijektiv afbildning af U^* på V. Denne afbildning har derfor en invers

$$B_1 : V \to U^*$$

givet ved

$$B_1(\mathbf{v}) = \mathbf{u}^* \quad \Leftrightarrow \quad A(u^*) = \mathbf{v}$$

Vi er nu i stand til at formulere definitionen på den pseudoinverse afbildning.



Figur 1.7: Skitse vedrørende pseudoinvers afbildning.

DEFINITION 1.1. Ved en **pseudoinvers** eller **generaliseret invers** afbildning til afbildningen A forstås en afbildning

$$B = B_1 \circ p_V : F \to E,$$

hvor p_V og B_1 er som nævnt foran. ▲

BEMÆRKNING 1.2. Den pseudoinverse er altså sammensætningen af projektionen ned på V langs V^* og den inverse til A's restriktion til U^*. ▼

BEMÆRKNING 1.3. Den pseudoinverse er selvfølgelig på ingen måde entydig, idet vi får en for hvert valg af underrummene U^* og V^*. ▼

Vi kan umiddelbart formulere nogle åbenbare egenskaber ved den pseudoinverse i følgende

SÆTNING 1.4. Den pseudoinverse B til A har følgende egenskaber

i) $\mathrm{rg}(B) = \mathrm{rg}(A) = s$

ii) $A \circ B = p_V : F \to V$

iii) $B \circ A = p_{U^*} : E \to U^*$

▲

Det kan vises, at disse egenskaber også karakteriserer pseudoinverse afbildninger, idet der gælder

SÆTNING 1.5. Lad $A : E \to F$ være lineær og med rang s. Antag, at B også har rang s, og at $A \circ B$ og $B \circ A$ begge er projektioner af rang s. Da er B en pseudoinvers af A som defineret ovenfor. ▲

BEVIS 1.4. Forbigås (relativt simpel øvelse i lineær algebra). ■

Vi vil nu give en matrixfremstilling af de ovenfor anførte definitioner.

DEFINITION 1.2. Lad $\mathbf{A}$ være en $(m \times n)$-matrix af rang s. En $(n \times m)$-matrix $\mathbf{B}$ af rang s, der tilfredsstiller

i) $\mathbf{A}\,\mathbf{B}$ idempotent med rang s

ii) $\mathbf{B}\,\mathbf{A}$idempotent med rang s,

kaldes en **pseudoinvers** eller **generaliseret** invers til $\mathbf{A}$. ▲

Ved hjælp af den pseudoinverse kan man karakterisere løsningsmængden til et system af lineære ligninger, idet vi har følgende

SÆTNING 1.6. Lad $\mathbf{A}$ og $\mathbf{B}$ være som i definition 1.2. Den generelle løsning til ligningen

$$\mathbf{A}\,\mathbf{x} = \mathbf{0}$$

er

$$(\mathbf{I} - \mathbf{B}\,\mathbf{A})\mathbf{z}, \qquad \mathbf{z} \in R^n,$$

og den generelle løsning til ligningen (**der forudsættes konsistent**)

$$\mathbf{A}\,\mathbf{x} = \mathbf{y},$$

er

$$\mathbf{B}\,\mathbf{y} + (\mathbf{I} - \mathbf{B}\,\mathbf{A})\mathbf{z}, \qquad \mathbf{z} \in R^n.$$

▲

BEVIS 1.5. Vi betragter først den homogene ligning. En løsning $\mathbf{x}$ er åbenbart et punkt i nulrummet $N(A) = A^{-1}(\mathbf{0})$ for den lineære afbildning svarende til $\mathbf{A}$. Matricen $\mathbf{B}\,\mathbf{A}$ svarer ifølge sætning 1.1 netop til projektionen ned på U^*. Følgelig svarer $\mathbf{I} - \mathbf{B}\,\mathbf{A}$ til projektionen ned på nulrummet $U = N(A)$, hvorfor et vilkårligt $\mathbf{x} \in N(A)$ kan skrives

$$\mathbf{x} = (\mathbf{I} - \mathbf{B}\,\mathbf{A})\mathbf{z}, \qquad \mathbf{z} \in R^n.$$

Dermed er påstanden angående den homogene ligning bevist.

Ligningen $\mathbf{A}\,\mathbf{x} = \mathbf{y}$ har kun løsning (i.e. er kun konsistent), hvis $\mathbf{y}$ ligger i billedrummet for $\mathbf{A}$. For et sådant $\mathbf{y}$ gælder imidlertid, at

$$\mathbf{A}\,\mathbf{B}\,\mathbf{y} = \mathbf{y},$$

ifølge sætning 1.4

Resultatet for den fuldstændige løsning følger umiddelbart. ■

For at illustrere begrebet anfører vi nu

EKSEMPEL 1.4. Vi betragter matricen

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 2 \\ 2 & 1 & 1 \\ 2 & 1 & 1 \end{bmatrix}.$$

$\mathbf{A}$ har åbenbart rangen 2.

Vi vil betragte den til $\mathbf{A}$ svarende lineære afbildning

$$A : E \to F$$

hvor E og F er 3-dimensionale vektorrum med baserne $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$ og $\{\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3\}$. Koordinaterne i disse baser benævnes med små x'er henholdsvis y'er, således at A har koordinatfremstillingen

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 2 \\ 2 & 1 & 1 \\ 2 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}.$$

Vi vil nu først bestemme **nulrummet**

$$U = N(A) = A^{-1}(\mathbf{0})$$

for A. Der gælder

$$\begin{aligned} \mathbf{x} \in U &\Leftrightarrow \mathbf{A}\,\mathbf{x} = \mathbf{0} \\ &\Leftrightarrow x_1 + x_2 + 2x_3 = 0 \quad \wedge \quad 2x_1 + x_2 + x_3 = 0 \\ &\Leftrightarrow x_1 = x_3 \quad \wedge \quad -3x_1 = x_2 \\ &\Leftrightarrow \mathbf{x}' = x_1(1, -3, 1). \end{aligned}$$

Nulrummet er altså

$$U = \{t \cdot \begin{bmatrix} 1 \\ -3 \\ 1 \end{bmatrix} | t \in R\} = \{t \cdot \mathbf{u}_3 | t \in R\}$$

Som **komplementært underrum** vælger vi at betragte det ortogonale komplement U^*. Dette har ligningen

$$(1, -3, 1)\mathbf{x} = 0,$$

eller

$$U^* = \{\mathbf{x} | x_1 - 3x_2 + x_3 = 0\}$$

Vi betragter nu en **ny basis** for E, nemlig $\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\}$. Koordinater i dette benævnes med små z'er. Overgangen fra z-koordinater til x-koordinater er givet ved

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 1 & 3 & 1 \\ 0 & 2 & -3 \\ -1 & 3 & 1 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix}$$

eller

$$\mathbf{x} = \mathbf{S}\mathbf{z}.$$

$\mathbf{S}$ matricens søjler er som bekendt $\mathbf{u}$'ernes koordinater i $\mathbf{e}$-systemet.

A's **billedrum** V er 2-dimensionalt og udspændes af $\mathbf{A}$'s søjler. Vi kan e.g. vælge de to første, nemlig

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} \quad , \quad \mathbf{v}_2 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

Som **komplementært underrum** V^* vælger vi V's ortogonale komplement. Det frembringes af krydsproduktet af $\mathbf{v}_1$ og $\mathbf{v}_2$:

$$\mathbf{v}_1 \times \mathbf{v}_2 = \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} = \mathbf{v}_3$$

Vi betragter nu den **nye basis** $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ for F. Koordinaterne heri benævnes med små w'er. Overgangen fra w-koordinater til y-koordinater er givet ved

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 2 & 1 & 1 \\ 2 & 1 & -1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix},$$

eller på kompakt form

$$\mathbf{y} = \mathbf{T}\,\mathbf{w}.$$

Vi vil nu bestemme koordinatudtryk for A i z- og w-koordinater. Da

$$\mathbf{y} = \mathbf{A}\,\mathbf{x}$$

fås

$$\mathbf{T}\,\mathbf{w} = \mathbf{A}\,\mathbf{S}\,\mathbf{z}$$

eller

$$\mathbf{w} = \mathbf{T}^{-1}\mathbf{A}\,\mathbf{S}\,\mathbf{z}.$$

Nu er

$$\mathbf{T}^{-1} = \begin{bmatrix} -1 & \frac{1}{2} & \frac{1}{2} \\ 2 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & \frac{1}{2} & -\frac{1}{2} \end{bmatrix},$$

hvorfor

$$\begin{aligned} \mathbf{T}^{-1}\mathbf{A}\,\mathbf{S} &= \begin{bmatrix} -1 & \frac{1}{2} & \frac{1}{2} \\ 2 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & \frac{1}{2} & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} 1 & 1 & 2 \\ 2 & 1 & 1 \\ 2 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 3 & 1 \\ 0 & 2 & -3 \\ -1 & 3 & 1 \end{bmatrix} \\ &= \begin{bmatrix} 2 & 0 & 0 \\ -3 & 11 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \end{aligned}$$

Da $\{\mathbf{u}_1, \mathbf{u}_2\}$ udspænder U^* og $\{\mathbf{v}_1, \mathbf{v}_2\}$ udspænder V, har restriktionen

$$A : U^* \to V$$

koordinatfremstillingen

$$\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ -3 & 11 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}.$$

Denne har den inverse afbildning

$$\begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 \\ \frac{3}{22} & \frac{1}{11} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}.$$

Opfattes punkterne som punkter i E og F og ikke blot som punkter i U^* og V fås

$$\begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ \frac{3}{22} & \frac{1}{11} & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} \tag{1.2}$$

Projektionen af F ned på V langs V^* har koordinatfremstillingen

$$\begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} \to \begin{bmatrix} w_1 \\ w_2 \\ 0 \end{bmatrix} \tag{1.3}$$

Dette er $z - w$ koordinatfremstillingen for den pseudoinverse B til afbildningen A. Vi ønsker imidlertid en beskrivelse i $x - y$ koordinater. Da

$$\mathbf{z} = \mathbf{S}^{-1}\mathbf{x} = \mathbf{C}\,\mathbf{w} = \mathbf{C}\,\mathbf{T}^{-1}\mathbf{y}$$

fås

$$\mathbf{x} = \mathbf{S}\,\mathbf{C}\,\mathbf{T}^{-1}\mathbf{y},$$

hvor $\mathbf{C}$ er matricen i ligning 1.1.

Vi har derfor

$$\begin{aligned} \mathbf{B} &= \mathbf{S}\,\mathbf{C}\,\mathbf{T}^{-1} \\ &= \begin{bmatrix} 1 & 3 & 1 \\ 0 & 2 & -3 \\ -1 & 3 & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{2} & 0 & 0 \\ \frac{3}{22} & \frac{1}{11} & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} -1 & \frac{1}{2} & \frac{1}{2} \\ 2 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & \frac{1}{2} & -\frac{1}{2} \end{bmatrix} \\ &= \frac{1}{22} \begin{bmatrix} -8 & 7 & 7 \\ 2 & 1 & 1 \\ 14 & -4 & -4 \end{bmatrix} \end{aligned}$$

Denne matrix er en pseudoinvers til $\mathbf{A}$. ♦

Som det fremgår af foranstående eksempel er det ret byrdefuldt blot at gå frem efter definitionen ved udledningen af en pseudoinvers. Ofte vil man kunne bruge følgende

SÆTNING 1.7. Lad $m \times n$ matricen $\mathbf{A}$ have rang s og lad

$$A = \begin{bmatrix} \mathbf{C} & \mathbf{D} \\ \mathbf{E} & \mathbf{F} \end{bmatrix},$$

hvor C er regulær og af dimension $s \times s$. En (mulig) pseudoinvers til A er da

$$\mathbf{A}^- = \begin{bmatrix} \mathbf{C}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},$$

hvor 0-matricerne har sådanne størrelser, at $\mathbf{A}^-$ har dimensionen $n \times m$. ▲

BEVIS 1.6. Vi har

$$\mathbf{A}\mathbf{A}^-\mathbf{A} = \begin{bmatrix} \mathbf{C} & \mathbf{D} \\ \mathbf{E} & \mathbf{F} \end{bmatrix} \begin{bmatrix} \mathbf{C}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{C} & \mathbf{D} \\ \mathbf{E} & \mathbf{F} \end{bmatrix} = \begin{bmatrix} \mathbf{C} & \mathbf{D} \\ \mathbf{E} & \mathbf{E}\mathbf{C}^{-1}\mathbf{D} \end{bmatrix}.$$

Da $\mathrm{rg}(\mathbf{A}) = s$, kan de sidste $n - s$ søjler skrives som linearkombinationer af de første s, d.v.s. der eksisterer en matrix H, så

$$\begin{bmatrix} \mathbf{D} \\ \mathbf{F} \end{bmatrix} = \begin{bmatrix} \mathbf{C} \\ \mathbf{E} \end{bmatrix} \mathbf{H}$$

eller

$$\begin{aligned} \mathbf{D} &= \mathbf{C}\mathbf{H} \\ \mathbf{F} &= \mathbf{E}\mathbf{H} \end{aligned}$$

Heraf fås

$$\mathbf{F} = \mathbf{E}\mathbf{C}^{-1}\mathbf{D}.$$

Indsættes dette i øverste formellinie, fås

$$\mathbf{A}\mathbf{A}^-\mathbf{A} = \mathbf{A}$$

Ved multiplikation til venstre med A^- henholdsvis til højre med A^- ses at A^-A og AA^- er idempotente. Sætningen følger nu af definitionen side 21 ■

Vi illustrerer anvendelsen af sætningen i nedenstående

EKSEMPEL 1.5. Vi betragter den i eksempel 1.4 givne matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 2 \\ 2 & 1 & 1 \\ 2 & 1 & 1 \end{bmatrix}.$$

Da

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

kan vi som pseudoinvers anvende

$$\mathbf{A}^{-} = \begin{bmatrix} -1 & 1 & 0 \\ 2 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

♦

Fordelen ved at anvende den i eksempel 1.4 beskrevne fremgangsmåde fremfor den langt enklere anførte i eksempel 1.5 er, at man der får en præcis geometrisk beskrivelse af forholdene.

BEMÆRKNING 1.4. Til slut må det anføres, at der i litteraturen optræder en række definitioner på pseudoinverse og generaliserede inverse, således at man ved en eventuel anvendelse må præcisere, hvad definitionen er. Her må især nævnes den såkaldte **Moore-Penrose** inverse $\mathbf{A}^{+}$ til en matrix $\mathbf{A}$. Den tilfredsstiller

i) $\mathbf{A}\,\mathbf{A}^{+}\mathbf{A} = \mathbf{A}$

ii) $\mathbf{A}^{+}\mathbf{A}\,\mathbf{A}^{+} = \mathbf{A}^{+}$

iii) $(\mathbf{A}\,\mathbf{A}^{+})' = \mathbf{A}\,\mathbf{A}^{+}$

iv) $(\mathbf{A}^{+}\mathbf{A})' = \mathbf{A}^{+}\mathbf{A}$

Det ses trivielt at en Moore-Penrose invers virkelig er en generaliseret invers. De øvrige betingelser sikrer at man ved mindste kvadraters løsning til en inkonsistent ligning opnår løsninger med minimal norm. Dette skal vi ikke komme ind på her, men blot henvise den interesserede læser til litteraturen, e.g. [29]. ▼

1.4 Egenværdiproblemer. Kvadratiske former

Vi indleder med de grundlæggende definitioner og sætninger i

1.4.1 Egenværdier og egenvektorer for symmetriske matricer

Selve den nedenfor anførte definition på en egenvektor og en egenværdi er gyldig for vilkårlige kvadratiske matricer, men vi vil overalt i det følgende forudsætte, at de involverede matricer er symmetriske med mindre andet explicit er nævnt.

En **egenværdi** λ for den symmetriske $n \times n$ matrix $\mathbf{A}$ er en løsning til ligningen

$$\det(\mathbf{A} - \lambda \mathbf{I}) = 0.$$

Der er n (reelle) egenværdier (nogle eventuelt sammenfaldende). Hvis λ er en egenværdi, eksisterer der vektorer $\mathbf{x} \neq 0$, således at

$$\mathbf{A}\,\mathbf{x} = \lambda \mathbf{x},$$

d.v.s. der findes vektorer, der ved den til $\mathbf{A}$ hørende lineære afbildning går over i et multiplum af sig selv. Sådanne vektorer kaldes **egenvektorer** svarende til egenværdien λ. Antallet af fra 0 forskellige egenværdier er lig $\mathrm{rg}(\mathbf{A})$. Her skal enhver egenværdi medtages så ofte som dens multiplicitet angiver. Mere interessant er

SÆTNING 1.8. Hvis λ_i og λ_j er forskellige egenværdier,og hvis $\mathbf{x}_i$ og $\mathbf{x}_j$ er tilhørende egenvektorer, da er $\mathbf{x}_i$ og $\mathbf{x}_j$ ortogonale, d.v.s. $\mathbf{x}_i'\mathbf{x}_j = 0$. ▲

BEVIS 1.7. Vi har

$$\begin{aligned} \mathbf{A}\,\mathbf{x}_i &= \lambda_i \mathbf{x}_i \\ \mathbf{A}\,\mathbf{x}_j &= \lambda_j \mathbf{x}_j \end{aligned}$$

Heraf fås umiddelbart

$$\begin{aligned} \mathbf{x}_j'\mathbf{A}\,\mathbf{x}_i &= \lambda_i \mathbf{x}_j'\mathbf{x}_i \\ \mathbf{x}_i'\mathbf{A}\,\mathbf{x}_j &= \lambda_j \mathbf{x}_i'\mathbf{x}_j. \end{aligned}$$

Vi transponerer den første relation og får

$$\mathbf{x}_i'\mathbf{A}'\mathbf{x}_j = \lambda_i \mathbf{x}_i'\mathbf{x}_j.$$

Da $\mathbf{A}$ er symmetrisk medfører dette, at

$$\lambda_i \mathbf{x}_i'\mathbf{x}_j = \lambda_j \mathbf{x}_i'\mathbf{x}_j,$$

og da $\lambda_i \neq \lambda_j$ er $\mathbf{x}_i'\mathbf{x}_j = 0$ d.v.s. $\mathbf{x}_i \perp \mathbf{x}_j$. ■

Resultatet i sætning 1.8 kan suppleres med følgende sætning, der nævnes uden bevis.

SÆTNING 1.9. Hvis λ er en egenværdi med multiplicitet m, da udgør mængden af egenvektorer svarende til λ et m-dimensionalt underrum. Dette indebærer specielt, at der eksisterer m ortogonale egenvektorer svarende til λ. ▲

Ved en kombinering af disse 2 sætninger bringes man let til at indse følgende

KOROLLAR 1.1. For en vilkårlig symmetrisk matrix $\mathbf{A}$ eksisterer der en basis for R^n bestående af indbyrdes ortogonale egenvektorer til $\mathbf{A}$.

Normeres vektorerne i en sådan basis bestående af ortogonale egenvektorer fås en **ortonormal** basis $(\mathbf{p}_1, \ldots, \mathbf{p}_n)$. Sætter vi $\mathbf{P}$ lig den $n \times n$ matrix, hvis søjler netop er koordinaterne til disse vektorer, d.v.s.

$$\mathbf{P} = (\mathbf{p}_1, \ldots, \mathbf{p}_n)$$

fås

$$\mathbf{P}'\mathbf{P} = \mathbf{I}$$

$\mathbf{P}$ er altså en **ortogonal** matrix, og

$$\mathbf{A}\,\mathbf{P} = \mathbf{P}\,\mathbf{\Lambda}$$

hvor $\mathbf{\Lambda}$ er en diagonalmatrix med egenværdierne for $\mathbf{A}$ (gentaget efter multiplicitet) i diagonalen. Ved hjælp af dette fås følgende

SÆTNING 1.10. Lad $\mathbf{A}$ være en symmetrisk matrix. Da eksisterer der en ortogonal matrix $\mathbf{P}$, således at

$$\mathbf{P}'\mathbf{A}\,\mathbf{P} = \mathbf{\Lambda}$$

hvor $\mathbf{\Lambda}$ er en diagonal matrix med $\mathbf{A}$'s egenværdier i diagonalen (gentaget så ofte multipliciteten angiver). Som $\mathbf{P}$ kan man vælge en matrix, hvis søjler er ortonormerede egenvektorer til $\mathbf{A}$. ▲

BEVIS 1.8. Triviel følge af ovenstående relation. ■

SÆTNING 1.11. Lad $\mathbf{A}$ være en symmetrisk matrix med ikke-negative egenværdier. Da eksisterer der en regulær matrix $\mathbf{B}$, således at

$$\mathbf{B}'\mathbf{A}\,\mathbf{B} = \mathbf{E},$$

hvor $\mathbf{E}$ er en diagonal matrix, der har 0'ler eller 1'ter i diagonalen. Antallet af 1'ter er lig $\mathrm{rg}(\mathbf{A})$. Hvis $\mathbf{A}$ har fuld rang, bliver $\mathbf{E}$ specielt en enhedsmatrix. ▲

BEVIS 1.9. Ved at multiplicere $\mathbf{P}$ (til højre) med en diagonal matrix $\mathbf{C}$, der i diagonalen har elementerne

$$c_i = \begin{cases} \frac{1}{\sqrt{\lambda_i}} & \text{hvis } \lambda_i > 0 \\ 1 & \text{hvis } \lambda_i = 0 \end{cases},$$

fås sætningen umiddelbart med $\mathbf{B} = \mathbf{P}\,\mathbf{C}$. ■

Relation i sætning 1.10 er ensbetydende med

$$\mathbf{A} = \mathbf{P}\,\mathbf{\Lambda}\,\mathbf{P}'$$

eller

$$\mathbf{A} = (\mathbf{p}_1 \ldots \mathbf{p}_n) \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} \mathbf{p}_1' \\ \vdots \\ \mathbf{p}_n' \end{bmatrix},$$

d.v.s. vi har følgende opspaltning af matricen

$$\mathbf{A} = \lambda_1 \mathbf{p}_1 \mathbf{p}_1' + \cdots + \lambda_n \mathbf{p}_n \mathbf{p}_n'.$$

Denne opspaltning af den symmetriske matrix $\mathbf{A}$ kaldes ofte dens **spektral fremstilling**, idet egenværdierne $\{\lambda_1, \ldots, \lambda_n\}$ betegnes matricens **spektrum**.

Med den åbenbare definition af $\mathbf{\Lambda}^{\frac{1}{2}}$, nemlig som $\mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n})$, ser vi, at vi kan skrive

$$\mathbf{A} = (\mathbf{P}\,\mathbf{\Lambda}^{\frac{1}{2}})(\mathbf{P}\,\mathbf{\Lambda}^{\frac{1}{2}})' = \mathbf{G}\,\mathbf{G}'.$$

Det kan her indskydes, at hvis $\mathbf{A}$ er positivt definit, findes der en fremstilling

$$\mathbf{A} = \mathbf{L}\,\mathbf{L}',$$

hvor $\mathbf{L}$ er en nedre trekantmatrix. Denne fremstilling kaldes **Cholesky faktoriseringen** af $\mathbf{A}$ (se f.eks. [34].

Endelig har vi

SÆTNING 1.12. Lad $\mathbf{A}$ være en regulær, symmetrisk matrix. Da har $\mathbf{A}$ og $\mathbf{A}^{-1}$ de samme egenvektorer, men svarende til reciprokke egenværdier. ▲

BEVIS 1.10. Lad λ være en egenværdi for $\mathbf{A}$ og $\mathbf{x}$ en tilhørende egenvektor, d.v.s.

$$\mathbf{A}\,\mathbf{x} = \lambda\mathbf{x}.$$

Da $\mathbf{A}$ er regulær, er dette ensbetydende med

$$\mathbf{A}^{-1}\mathbf{x} = \frac{1}{\lambda}\mathbf{x},$$

hvilket skulle vises. ■

Endelig bemærkes, at

$$\det \mathbf{A} = \prod_i \lambda_i.$$

EKSEMPEL 1.6. **Ortogonale transformationer af planen.** For at give en geometrisk forståelse af de transformationer, der reducerer en symmetrisk matrix til diagonalform, gennemgår vi de ortogonale transformationer af planen.

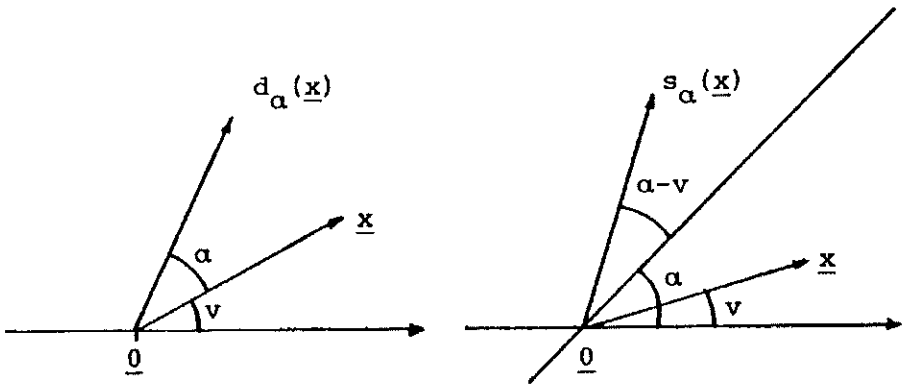
Ved at udnytte ortogonalitetsbetingelserne $\mathbf{P}'\mathbf{P} = \mathbf{I}$ ses let, at de eneste ortogonale 2×2-matricer er matricer af formen

$$\begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix} \quad \text{og} \quad \begin{bmatrix} \cos\alpha & \sin\alpha \\ \sin\alpha & -\cos\alpha \end{bmatrix}.$$

Vi vil nu vise, at disse svarer til **drejninger** (rotationer) om origo og **spejlinger** (refleksioner) i rette linier.

Vi gør det ved at bestemme koordinatudtryk for de lineære afbildninger d_α og s_α, der henholdsvis repræsenterer en drejning på vinklen α af planen og en spejling i linien, der danner vinklen α med 1.ste aksen.

Afbildningerne er skitseret i figur 1.6. Idet $\mathbf{x} = r(\cos v, \sin v)'$, hvor r kan sættes lig



Figur 1.8: Drejning og spejling bestemt af vinklen α.

1, har vi

$$d_\alpha(\mathbf{x}) = \begin{bmatrix} \cos(\alpha+v) \\ \sin(\alpha+v) \end{bmatrix} = \begin{bmatrix} \cos\alpha\cos v - \sin\alpha\sin v \\ \sin\alpha\cos v + \cos\alpha\sin v \end{bmatrix}$$
$$= \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} \cos v \\ \sin v \end{bmatrix}.$$

Heraf ses, at d_α **har matrix fremstillingen**

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \to \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

Analogt fås

$$s_\alpha(\mathbf{x}) = \begin{bmatrix} \cos(2\alpha - v) \\ \sin(2\alpha - v) \end{bmatrix} = \begin{bmatrix} \cos 2\alpha\cos v + \sin 2\alpha\sin v \\ \sin 2\alpha\cos v - \cos 2\alpha\sin v \end{bmatrix}$$
$$= \begin{bmatrix} \cos 2\alpha & \sin 2\alpha \\ \sin 2\alpha & -\cos 2\alpha \end{bmatrix} \begin{bmatrix} \cos v \\ \sin v \end{bmatrix}.$$

hvorfor s_α **har matrix fremstillingen**

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \to \begin{bmatrix} \cos 2\alpha & \sin 2\alpha \\ \sin 2\alpha & -\cos 2\alpha \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

Hermed er påstanden i indledningen bevist.

Det kan endvidere være nyttigt at erindre sig følgende relationer mellem drejninger og spejlinqer af planen

$$s_{\frac{\pi}{4}} \circ d_\alpha = s_{\frac{\pi}{4}-\frac{\alpha}{2}}$$
$$s_\alpha = s_{\frac{\pi}{4}} \circ d_{\frac{\pi}{2}-2\alpha}.$$

Den første relation følger af

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix} =$$
$$\begin{bmatrix} \sin\alpha & \cos\alpha \\ \cos\alpha & \sin\alpha \end{bmatrix} = \begin{bmatrix} \cos(\frac{\pi}{4}-\alpha) & \sin(\frac{\pi}{4}-\alpha) \\ \sin(\frac{\pi}{4}-\alpha) & -\cos(\frac{\pi}{4}-\alpha) \end{bmatrix}.$$

Den sidste af de to relationer fås af den første ved at erstatte α med $\frac{\pi}{2} - 2\alpha$. ♦

I det næste afsnit skal vi bl.a. betragte problemet med at generalisere spektralfremstillingen til en vilkårlig matrix.

1.4.2 Singulær-værdi dekomposition af vilkårlig matrix. Q- og R-modus-analyser

Vi giver først hovedresultatet, der også kendes under navnet **Eckart-Young's sætning**.

SÆTNING 1.13. Lad $\mathbf{x}$ være en vilkårlig $n \times p$ matrix med rang r. Da eksisterer der ortogonale matricer $\mathbf{U}$ $(p \times r)$ og $\mathbf{V}$ $(n \times r)$ samt positive tal $\gamma_1, \ldots, \gamma_r$, således at

$$x = \mathbf{V}\boldsymbol{\Gamma}\mathbf{U}' = [\mathbf{v}_1 \cdots \mathbf{v}_r] \begin{bmatrix} \gamma_1 \cdots 0 \\ \vdots \quad \vdots \\ 0 \cdots \gamma_r \end{bmatrix} \begin{bmatrix} \mathbf{u}_1' \\ \vdots \\ \mathbf{u}_r' \end{bmatrix} = \gamma_1 \mathbf{v}_1 \mathbf{u}_1' + \cdots + \gamma_r \mathbf{v}_r \mathbf{u}_r',$$

hvor $\boldsymbol{\Gamma} = \text{diag}(\gamma_1, \ldots, \gamma_r)$ og $\mathbf{v}_1, \ldots, \mathbf{v}_r$ er søjlerne i $\mathbf{V}$ og $\mathbf{u}_1, \ldots, \mathbf{u}_r$ søjlerne i $\mathbf{U}$. ▲

BEVIS 1.11. Forbigås. Se f.eks. [17]. ■

Tallene $\gamma_1, \ldots, \gamma_r$ kaldes $\mathbf{x}$'s **singulære værdier**.

I det følgende vil vi nu undersøge sammenhængen mellem $\mathbf{x}$'s singulære værdier og egenværdiproblemerne for de symmetriske matricer $\mathbf{x}\,\mathbf{x}'$ $(n \times n)$ og $\mathbf{x}'\mathbf{x}$ $(p \times p)$.

Vi anfører dog først

SÆTNING 1.14. For en vilkårlig (reel) matrix $\mathbf{x}$ gælder, at $\mathbf{x}'\mathbf{x}$ og $\mathbf{x}\,\mathbf{x}'$ har ikke-negative egenværdier og

$$\mathrm{rg}(\mathbf{x}'\mathbf{x}) = \mathrm{rg}(\mathbf{x}\,\mathbf{x}') = \mathrm{rg}(\mathbf{x})$$

▲

BEVIS 1.12. Det er tilstrækkeligt at vise resultaterne for $\mathbf{x}'\mathbf{x}$. Det er åbenbart, at $\mathbf{x}'\mathbf{x}$ er symmetrisk, hvorfor der eksisterer en ortogonal matrix $\mathbf{P}$, så

$$\mathbf{P}'\mathbf{x}'\mathbf{x}\,\mathbf{P} = \mathbf{\Lambda}$$

d.v.s.

$$(\mathbf{x}\,\mathbf{P})'(\mathbf{x}\mathbf{P}) = \mathbf{\Lambda}.$$

Sættes $\mathbf{x}\,\mathbf{P} = \mathbf{B} = (b_{ij})$, fås $\mathbf{B}'\mathbf{B} = \mathbf{\Lambda}$, d.v.s.

$$\lambda_i = \sum_j b_{ij}^2 > 0,$$

d.v.s. $\mathbf{x}'\mathbf{x}$ har ikke-negative egenvektorer. Det ses endvidere, at

$$\begin{aligned} \mathrm{rg}(\mathbf{x}'\mathbf{x}) &= \mathrm{card}(\lambda_i \neq 0) \\ &= \mathrm{card}\{\text{søjler } \mathbf{b}_j \text{ i } \mathbf{B}, \text{ der er} \neq \mathbf{0}\} \end{aligned}$$

Da $\mathbf{b}_i'\mathbf{b}_j = 0$ for $i \neq j$ (ifølge ligning 1.1) er derfor

$$\mathrm{rg}(\mathbf{x}'\mathbf{x}) = \mathrm{rg}(\mathbf{B})$$

Da $\mathbf{P}$ er regulær, får vi af et resultat side 13, at

$$\mathrm{rg}(\mathbf{B}) = \mathrm{rg}(\mathbf{x}\,\mathbf{P}) = \mathrm{rg}(\mathbf{x}).$$

■

Vi anfører et lille corollar til sætningen.

KOROLLAR 1.2. Lad Σ være symmetrisk og positivt definit. Da er for en vilkårlig matrix $\mathbf{x}$

$$\mathrm{rg}(\mathbf{x}'\Sigma^{-1}\mathbf{x}) = \mathrm{rg}(\mathbf{x}),$$

naturligvis forudsat, at de involverede produkter eksisterer.

BEVIS 1.13. Da Σ^{-1} også er regulær og positivt definit, eksisterer der en ortogonal matrix $\mathbf{P}$, så

$$\mathbf{P}'\Sigma^{-1}\mathbf{P} = \mathbf{\Lambda},$$

hvor $\mathbf{\Lambda}$ er en diagonalmatrix. Dette medfører

$$\Sigma^{-1} = \mathbf{P}\,\mathbf{\Lambda}\,\mathbf{P}' = \mathbf{P}\,\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{\Lambda}^{\frac{1}{2}}\mathbf{P}' = \mathbf{P}\,\mathbf{\Lambda}^{\frac{1}{2}}(\mathbf{P}\,\mathbf{\Lambda}^{\frac{1}{2}})' = \mathbf{B}\,\mathbf{B}'.$$

Her betegner $\mathbf{\Lambda}^{\frac{1}{2}}$ den diagonalmatrix, hvis diagonal elementer er kvadratroden af de tilsvarende elementer i $\mathbf{\Lambda}$. Det er trivielt, at $\mathbf{B}$ er regulær. Denne relation indsættes, og vi får

$$\mathbf{x}'\Sigma^{-1}\mathbf{x} = \mathbf{x}'\mathbf{B}\,\mathbf{B}'\mathbf{x} = (\mathbf{B}'\mathbf{x})'\mathbf{B}'\mathbf{x},$$

d.v.s.

$$\mathrm{rg}(\mathbf{x}'\Sigma^{-1}\mathbf{x}) = \mathrm{rg}(\mathbf{B}'\mathbf{x}) = \mathrm{rg}(\mathbf{x}),$$

hvilket skulle vises. ∎

Vi har nu med betegnelserne fra sætning 1.14.

SÆTNING 1.15.

i) matricen $\mathbf{x}\,\mathbf{x}'$ $(n \times n)$ har r positive egenværdier og $n - r$ egenværdier lig 0. De positive egenværdier er $\gamma_1^2, \ldots, \gamma_r^2$, hvor $\gamma_1, \ldots, \gamma_r$ er de singulære værdier for $\mathbf{x}$. De tilsvarende egenvektorer er $\mathbf{v}_1, \ldots, \mathbf{v}_r$.

ii) Tilsvarende har $\mathbf{x}'\mathbf{x}$ $(p \times p)$ r positive og $(p - r)$ 0-egenværdier. De positive egenværdier er $\gamma_1^2, \ldots, \gamma_r^2$ og de tilsvarende egenvektorer er $\mathbf{u}_1, \ldots, \mathbf{u}_r$.

iii) De positive egenværdier for $\mathbf{x}\,\mathbf{x}'$ og $\mathbf{x}'\mathbf{x}$ er altså ens og relationen mellem de tilsvarende egenvektorer er $(m = 1, \ldots, r)$

$$\mathbf{v}_m = \frac{1}{\gamma_m}\mathbf{x}\,\mathbf{u}_m \quad \text{og} \quad \mathbf{u}_m = \frac{1}{\gamma_m}\mathbf{x}'\mathbf{v}_m,$$

eller på kompakt form

$$\mathbf{V} = \mathbf{x}\,\mathbf{U}\,\boldsymbol{\Gamma}^{-1} \quad \text{og} \quad \mathbf{U} = \mathbf{x}'\mathbf{V}\,\boldsymbol{\Gamma}^{-1}$$

▲

BEVIS 1.14. Følger ved anvendelse af Eckart-Young's sætning. ■

BEMÆRKNING 1.5. Analysen af matricen $\mathbf{x}'\mathbf{x}$ kaldes ***R*-modus analyse** og analysen af $\mathbf{x}\,\mathbf{x}'$ ***Q*-modus analyse**, betegnelser, der stammer fra faktoranalysen, jvf. kapitel 8. ▼

BEMÆRKNING 1.6. Det fremgår af sætningen, at man kan opnå resultaterne fra en R-modus analyse ud fra en Q-modus analyse og omvendt. Ved en konkret anvendelse bør man derfor betragte den af matricerne $\mathbf{x}'\mathbf{x}$ og $\mathbf{x}\,\mathbf{x}'$, der er af mindst orden. ▼

1.4.3 Kvadratiske former og positivt definitte matricer

I dette afsnit betragter vi fremdeles kun symmetriske matricer.

Ved den **kvadratiske form** svarende til den symmetriske matrix $\mathbf{A}$ forstås afbildningen

$$\mathbf{x} \to \mathbf{x}'\mathbf{A}\,\mathbf{x} = \sum a_{ii}x_i^2 + 2\sum_{1<j} a_{ij}x_i x_j.$$

Vi siger, at en symmetrisk matrix $\mathbf{A}$ er **positivt definit**, **respektive positivt semidefinit**, hvis den tilhørende kvadratiske form er positiv, respektive ikke negativ, i vektorer forskellige fra 0-vektoren, i.e. såfremt

$$\forall \mathbf{x} \neq \mathbf{0} : \mathbf{x}'\mathbf{A}\,\mathbf{x} > 0,$$

respektive

$$\forall \mathbf{x} \neq 0 : \mathbf{x}'\mathbf{A}\,\mathbf{x} \geq 0.$$

Vi siger da også, at den kvadratiske form er **positivt definit**, **respektive positivt semidefinit**.

Vi har følgende

SÆTNING 1.16. Den symmetriske matrix $\mathbf{A}$ er positivt definit, respektive semidefinit, hvis alle $\mathbf{A}$'s egenværdier er positive, respektive ikke negative. ▲

BEVIS 1.15. Vi har med $\mathbf{P}$ som i sætning 1.10

$$\begin{aligned}\mathbf{x}'\mathbf{A}\mathbf{x} &= \mathbf{x}'\mathbf{P}'\mathbf{P}\,\mathbf{A}\,\mathbf{P}\,\mathbf{P}'\mathbf{x} = (\mathbf{P}'\mathbf{x})'\mathbf{\Lambda}(\mathbf{P}'\mathbf{x}) \\ &= \mathbf{y}'\mathbf{\Lambda}\,\mathbf{y} = \lambda_1 y_1^2 + \cdots + \lambda_n y_n^2.\end{aligned}$$

■

Et andet nyttigt resultat er

SÆTNING 1.17. En symmetrisk $n \times n$ matrix $\mathbf{A}$ er positivt definit, netop hvis alle hovedunderdeterminanter

$$d_i = \det \begin{bmatrix} a_{11} & \cdots & a_{1i} \\ \vdots & & \vdots \\ a_{i1} & \cdots & a_{ii} \end{bmatrix}, \qquad i = 1, \ldots, n,$$

er positive. ▲

BEVIS 1.16. Forbigås ■

Vi anfører nu en meget vigtig sætning om ekstrema af kvadratiske former

SÆTNING 1.18. Sættes egenværdierne for den symmetriske matrix $\mathbf{A}$ lig $\lambda_1 \geq \cdots \geq \lambda_n$ med tilhørende egenvektorer $\mathbf{p}_1, \ldots, \mathbf{p}_n$, og defineres

$$R(\mathbf{x}) = \frac{\mathbf{x}'\mathbf{A}\,\mathbf{x}}{\mathbf{x}'\mathbf{x}},$$

samt

$$M_k = \{\mathbf{x} | \mathbf{x}'\mathbf{p}_i = 0, \qquad i = 1, \ldots, k-1\},$$

Da gælder, at

$$\begin{aligned} \sup_{\mathbf{x}} R(\mathbf{x}) &= R(\mathbf{p}_1) = \lambda_1, \\ \inf_{\mathbf{x}} R(\mathbf{x}) &= R(\mathbf{p}_n) = \lambda_n, \\ \sup_{\mathbf{x}\in M_k} R(\mathbf{x}) &= R(\mathbf{p}_k) = \lambda_k. \end{aligned}$$

▲

BEVIS 1.17. En vilkårlig vektor $\mathbf{x}$ kan skrives

$$\mathbf{x} = \alpha_1\mathbf{p}_1 + \cdots + \alpha_n\mathbf{p}_n.$$

Hvis $\mathbf{p}_i'\mathbf{x} = 0$, $i = 1, \ldots, k-1$, fås $\alpha_1 = \cdots = \alpha_{k-1} = 0$, d.v.s.

$$\mathbf{x} = \alpha_k\mathbf{p}_k + \cdots + \alpha_n\mathbf{p}_n.$$

Følgelig er

$$\mathbf{x}'\mathbf{A}\,\mathbf{x} = \alpha_k^2\lambda_k + \cdots + \alpha_n^2\lambda_n,$$

og

$$R(\mathbf{x}) = \frac{\mathbf{x}'\mathbf{A}\,\mathbf{x}}{\mathbf{x}'\mathbf{x}} = \frac{\alpha_k^2\lambda_k + \cdots + \alpha_n^2\lambda_n}{\alpha_k^2 + \cdots + \alpha_n^2}$$

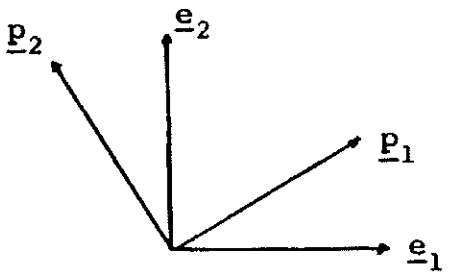
Det er klart, at dette udtryk er maksimalt for

$$(\alpha_k, \ldots, \alpha_n) = (\alpha_k, 0, \ldots, 0),$$

hvor det antager værdien λ_k. Resultatet med inf vises analogt. ■

BEMÆRKNING 1.7. Sætningen giver for $k = 1$, at den enhedsvektor, i.e. den "retning", hvor den kvadratiske form antager sin største værdi, netop er egenvektoren svarende til den største egenværdi. Betragtes den kvadratiske form kun i enhedsvektorer, der er ortogonale på egenvektorer svarende til de $k-1$ største egenværdier siger sætningen, at maksimum fås i den retning, der svarer til egenvektoren hørende til den k'te største egenværdi. ▼

BEMÆRKNING 1.8. R(x) kaldes også **Rayleigh's kvotient**. ▼



Figur 1.9: Skitse visende basisskift.

Vi vil nu beskrive niveaukurverne for positivt definitte kvadratiske former.

SÆTNING 1.19. Lad $\mathbf{A}$ være en positiv definit. Da er løsningsmængden til ligningen

$$\mathbf{x}'\mathbf{A}\,\mathbf{x} = c, \qquad c > 0,$$

en ellipsoide med hovedakser i egenvektorernes retninger. Første hovedakse svarer til den mindste egenværdi, anden til den næstmindste etc. ▲

BEVIS 1.18. Vi betragter matricen $\mathbf{P} = (\mathbf{p}_1, \ldots, \mathbf{p}_n)$, hvis søjler er koordinaterne til ortonormerede egenvektorer til $\mathbf{A}$. Da gælder der med $\mathbf{y} = \mathbf{P}'\mathbf{x}$

$$\begin{aligned} \mathbf{x}'\mathbf{A}\,\mathbf{x} &= \mathbf{y}'\Lambda\mathbf{y} \\ &= \lambda_1 y_1^2 + \cdots + \lambda_n y_n^2 \\ &= \frac{y_1^2}{(1/\sqrt{\lambda_1})^2} + \cdots + \frac{y_n^2}{(1/\sqrt{\lambda_n})^2} \end{aligned} \tag{1.4}$$

matrixligningen

$$\mathbf{y} = \mathbf{P}'\mathbf{x} \quad \Leftrightarrow \quad \mathbf{x} = \mathbf{P}\,\mathbf{y}$$

svarer til et basisskifte fra den oprindelige ortonormale basis $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ til den ortonormale basis $\{\mathbf{p}_1, \ldots, \mathbf{p}_n\}$.

Lad nemlig S være et punkt, hvis $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$-koordinater kaldes $\mathbf{x}$ og hvis $\{\mathbf{p}_1, \ldots, \mathbf{p}_n\}$-koordinater kaldes $\mathbf{y}$. Da gælder

$$x_1\mathbf{e}_1 + \cdots + x_n\mathbf{e}_n = y_1\mathbf{p}_1 + \cdots + y_n\mathbf{p}_n,$$

eller

$$(\mathbf{e}_1 \cdots \mathbf{e}_n)\mathbf{x} = (\mathbf{p}_1 \cdots \mathbf{p}_n)\mathbf{y},$$

d.v.s.

$$\mathbf{I}\,\mathbf{x} = \mathbf{P}\,\mathbf{y},$$

hvor $\mathbf{I}$ er en enhedsmatrix.

Udtrykket i 1.4 viser derfor løsningsmængdens ligning i y-koordinater svarende til koordinatsystemet bestående af ortonormerede egenvektorer. Dette viser, at der er tale om en ellipsoide. Resten af sætningen følger nu ved at bemærke, at 1.ste hovedakse svarer til det y_i, for hvilken $1/\sqrt{\lambda_i}$ er maksimal, d.v.s. for hvilken λ_i er minimal. ■

BEMÆRKNING 1.9. Hvis matricen kun er positiv semidefinit svarer løsningsmængden til ligningen til en **elliptisk cylinder**. Foretager vi nemlig et basisskifte til basen $\{\mathbf{p}_1, \ldots, \mathbf{p}_n\}$ bestående af ortonormale egenvektorer, hvor vi for simpelheds skyld antager, at $\mathbf{p}_1, \ldots, \mathbf{p}_r$ svarer til de fra 0 forskellige egenværdier, får vi

$$\begin{aligned} \mathbf{x}'\mathbf{A}\,\mathbf{x} = c \quad &\Leftrightarrow \quad \lambda_1 y_1^2 + \cdots + \lambda_r y_r^2 + 0y_{r+1}^2 + \cdots + 0y_n^2 = c \\ &\Leftrightarrow \quad \lambda_1 y_1^2 + \cdots + \lambda_r y_r^2 = c. \end{aligned}$$

Heraf følger påstanden. Betragter vi den kvadratiske forms restriktion til underrummet udspændt af egenvektorerne svarende til egenværdier > 0, bliver løsningsmængden en ellipsoide. ▼

EKSEMPEL 1.7. Vi betragter den symmetriske positivt definitte matrix

$$\mathbf{A} = \begin{bmatrix} 3 & \sqrt{2} \\ \sqrt{2} & 2 \end{bmatrix}.$$

Den til $\mathbf{A}$ svarende kvadratiske form er

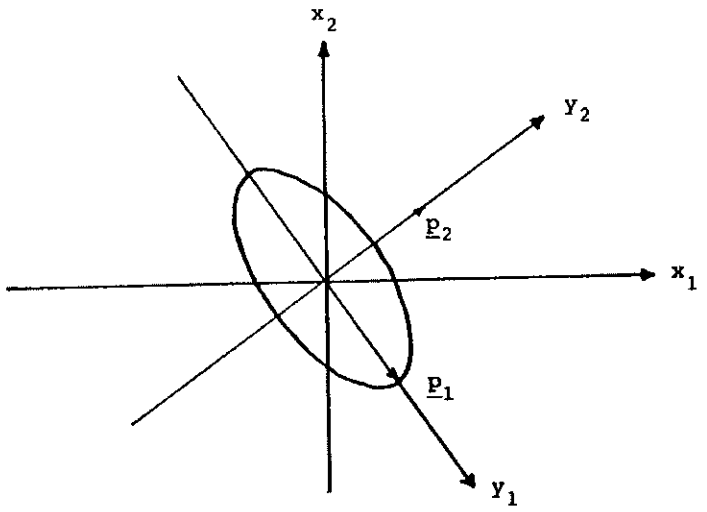
$$\mathbf{x}'\mathbf{A}\,\mathbf{x} = 3x_1^2 + 2x_2^2 + 2\sqrt{2}x_1x_2,$$

således at enhedsellipsen svarende til $\mathbf{A}$ er løsningsmængden til ligningen

$$3x_1^2 + 2x_2^2 + 2\sqrt{2}x_1x_2 = 1.$$

For at bestemme hovedakserne finder vi $\mathbf{A}$'s egenværdier. Vi finder

$$\begin{aligned} \det(\mathbf{A} - \lambda\,\mathbf{I}) = 0 \quad &\Leftrightarrow \quad \lambda^2 - 5\lambda + 4 = 0 \\ &\Leftrightarrow \quad \lambda = 1 \quad \vee \quad \lambda = 4. \end{aligned}$$



Figur 1.10: Ellipse bestemt af den i eksempel 1.7 anførte kvadratiske form.

Egenvektorer svarende til $\lambda = 1$ henholdsvis $\lambda = 4$ ses at være af formen $t(1, -\sqrt{2})$ henholdsvis $t(1, \sqrt{2}/2)$. Normeres disse fås

$$\mathbf{p}_1 = \begin{bmatrix} \frac{\sqrt{3}}{3} \\ \frac{-\sqrt{6}}{3} \end{bmatrix} \quad , \quad \mathbf{p}_2 = \begin{bmatrix} \frac{\sqrt{6}}{3} \\ \frac{-\sqrt{3}}{3} \end{bmatrix}.$$

Vælges basen $\{\mathbf{p}_1, \mathbf{p}_2\}$, bliver koordinatfremstillingen for den kvadratiske form

$$\mathbf{y} \to y_1^2 + 4y_2^2,$$

Ellipsen får ligningen

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

Den er skitseret i figur 1.7.

Da

$$\begin{aligned} \mathbf{p}_1 &= \begin{bmatrix} \frac{\sqrt{3}}{3} \\ \frac{-\sqrt{6}}{3} \end{bmatrix} = \begin{bmatrix} 0.577 \\ -0.820 \end{bmatrix} \\ &\simeq \begin{bmatrix} \cos(-54.7°) \\ \sin(-54,7°) \end{bmatrix}, \end{aligned}$$

svarer det nye koordinatsystem til en drejning af det gamle med vinklen $-54.7°$. ♦

1.4.4 Det generelle egenværdiproblem for symmetriske matricer

I forbindelse med teorien for kanoniske korrelationer og i diskriminantanalysen får vi brug for et lidt mere generelt egenværdibegreb end det hidtil betragtede. Vi indfører begrebet i

DEFINITION 1.3. Lad $\mathbf{A}$ og $\mathbf{B}$ være reelle $m \times m$ symmetriske matricer og lad $\mathbf{B}$ være af fuld rang. Et tal λ, for hvilket

$$\det(\mathbf{A} - \lambda\mathbf{B}) = 0,$$

benævnes en **egenværdi af A m.h.t. B**. For et sådant λ findes $\mathbf{x} \neq 0$ så

$$\mathbf{A}\,\mathbf{x} = \lambda\mathbf{B}\,\mathbf{x}.$$

En sådan vektor $\mathbf{x}$ kaldes en **egenvektor for A m.h.t. B**. ▲

BEMÆRKNING 1.10. De anførte begreber føres umiddelbart tilbage til egenværdi og egenvektor for den **ikke**-symmetriske matrix $\mathbf{B}^{-1}\mathbf{A}$. ▼

SÆTNING 1.20. Vi betragter igen situationen i definition 1.3, og lad yderligere $\mathbf{B}$ være positivt definit. Der findes da m reelle egenværdier af $\mathbf{A}$ m.h.t. $\mathbf{B}$. Hvis $\mathbf{A}$ er positivt semidefinit vil disse være ikke-negative, og hvis $\mathbf{A}$ er positivt definit, vil de være positive. ▲

BEVIS 1.19. Ifølge sætning 1.11 findes en matrix $\mathbf{I}$ med

$$\mathbf{T}'\mathbf{B}\,\mathbf{T} = \mathbf{I}.$$

Vi sætter

$$\mathbf{D} = \mathbf{T}'\mathbf{A}\,\mathbf{T}$$

$\mathbf{D}$ er åbenbart symmetrisk, og da

$$\mathbf{x}'\mathbf{D}\,\mathbf{x} = (\mathbf{T}\,\mathbf{x})'\mathbf{A}(\mathbf{T}\,\mathbf{x}),$$

ses at $\mathbf{D}$ og $\mathbf{A}$ er positivt semidefinitte, resp. positivt definitte samtidigt.

Nu er

$$\begin{aligned}(\mathbf{D}-\lambda\mathbf{I})\mathbf{v}=0 &\Leftrightarrow (\mathbf{T}'\mathbf{A}\,\mathbf{T}-\lambda\mathbf{T}'\mathbf{B}\,\mathbf{T})\mathbf{v}=0\\ &\Leftrightarrow (\mathbf{A}-\lambda\mathbf{B})(\mathbf{T}\,\mathbf{v})=0\end{aligned}$$

Heraf følger,at $\mathbf{D}$'s egenværdier er lig $\mathbf{A}$'s egenværdier m.h.t. $\mathbf{B}$, og at egenvektorerne for $\mathbf{A}$ m.h.t. $\mathbf{B}$ fås ved at anvende transformationen $\mathbf{T}$ på $\mathbf{D}$'s egenvektorer. Resultatet vedrørende egenværdiernes fortegn følger trivielt. ■

SÆTNING 1.21. Lad situationen være som ovenfor. Da eksisterer en basis for R^m bestående af egenvektorer $\mathbf{u}_1,\ldots,\mathbf{u}_m$ af $\mathbf{A}$ m.h.t. $\mathbf{B}$. Disse vektorer kan vælges konjugerede såvel m.h.t. $\mathbf{A}$ som m.h.t. $\mathbf{B}$, d.v.s.

$$\mathbf{u}_i'\mathbf{A}\,\mathbf{u}_j=\mathbf{u}_i'\mathbf{B}\,\mathbf{u}_j=0.$$

▲

BEVIS 1.20. Følger af beviset for ovenstående sætning og af corollar til sætning 1.9, når det erindres, at

$$0=\mathbf{v}_i'\mathbf{v}_j=(\mathbf{v}_i'\mathbf{T}')\mathbf{T}'^{-1}\mathbf{T}^{-1}(\mathbf{T}\,\mathbf{v}_j)=\mathbf{u}_i'\mathbf{B}\,\mathbf{u}_j,$$

hvor $\mathbf{v}_i,\ldots,\mathbf{v}_m$ er en ortonormal basis for R^m bestående af egenvektorer til $\mathbf{D}$.

Endelig er

$$\mathbf{u}_i'\mathbf{A}\,\mathbf{u}_j=\lambda_j\mathbf{u}_i'\mathbf{B}\,\mathbf{u}_j=0$$

■

SÆTNING 1.22. Lad $\mathbf{A}$ være symmetrisk og lad $\mathbf{B}$ være positivt definit. Da eksisterer en regulær matrix $\mathbf{R}$ med

$$\mathbf{R}'\mathbf{A}\,\mathbf{R}=\mathbf{\Lambda}=\operatorname{diag}(\lambda_1,\ldots,\lambda_n),$$

og

$$\mathbf{R}'\mathbf{B}\,\mathbf{R}=\mathbf{I},$$

hvor $\lambda_1,\ldots,\lambda_n$ er egenværdierne for $\mathbf{A}$ m.h.t. $\mathbf{B}$. Benævnes den i'te søjle i $\mathbf{R}'^{-1}$ s_i kan disse relationer skrives

$$\mathbf{A}=\lambda_1\mathbf{s}_1\mathbf{s}_1'+\cdots+\lambda_m\mathbf{s}_m\mathbf{s}_m',$$

og

$$\mathbf{B} = \mathbf{s}_1\mathbf{s}_1' + \ldots + \mathbf{s}_m\mathbf{s}_m'.$$

▲

BEVIS 1.21. Vi betragter det i beviset for sætning 1.20 anførte $\mathbf{D} = \mathbf{T}'\mathbf{A}\,\mathbf{T}$. Da $\mathbf{D}$ er symmetrisk, findes ifølge sætning 1.10 en ortogonal matrix $\mathbf{C}$ med

$$\mathbf{C}'\mathbf{D}\,\mathbf{C} = \mathbf{\Lambda},$$

idet vi erindrer, at $\mathbf{D}$'s egenværdier netop er $\mathbf{A}$'s egenværdier m.h.t. $\mathbf{B}$.

Vælges $\mathbf{R} = \mathbf{T}\,\mathbf{C}$, fås

$$\mathbf{R}'\mathbf{B}\,\mathbf{R} = \mathbf{C}'\mathbf{T}'\mathbf{B}\,\mathbf{T}\,\mathbf{C} = \mathbf{C}'\mathbf{C} = \mathbf{I},$$

og

$$\mathbf{R}'\mathbf{A}\,\mathbf{R} = \mathbf{C}'\mathbf{T}'\mathbf{A}\,\mathbf{T}\,\mathbf{C} = \mathbf{C}'\,\mathbf{D}\,\mathbf{C} = \mathbf{\Lambda}.$$

■

Endelig anfører vi en analog til sætning 1.18, nemlig

SÆTNING 1.23. Lad $\mathbf{A}$ være positivt semidefinit og $\mathbf{B}$ positivt definit. Lad $\mathbf{A}$'s egenværder m.h.t. $\mathbf{B}$ være $\lambda_1 \geq \cdots \geq \lambda_m$ og lad $\mathbf{v}_1, \ldots, \mathbf{v}_m$ betegne en basis for R^m bestående af de tilsvarende egenvektorer med $\mathbf{v}_i\mathbf{B}\,\mathbf{v}_j = 0 \quad i \neq j$. Vi sætter

$$R(\mathbf{x}) = \frac{\mathbf{x}'\mathbf{A}\,\mathbf{x}}{\mathbf{x}'\mathbf{B}\,\mathbf{x}}$$

samt

$$M_k = \{\mathbf{x} | \mathbf{x}'\mathbf{B}\,\mathbf{v}_1 = \cdots = \mathbf{x}'\mathbf{B}\,\mathbf{v}_{k-1} = 0\},$$

og har da

$$\begin{aligned}
\sup_{x} R(\mathbf{x}) &= R(\mathbf{v}_1) = \lambda_1 \\
\inf_{x} R(\mathbf{x}) &= R(\mathbf{v}_m) = \lambda_m \\
\sup_{\mathbf{x}\in M_k} R(\mathbf{x}) &= R(\mathbf{v}_k) = \lambda_k.
\end{aligned}$$

▲

BEVIS 1.22. Uden tab af generalitet kan $\mathbf{v}_i$'erne vælges så $\mathbf{v}_i'\mathbf{B}\mathbf{v}_i = 1$, og da en vilkårlig vektor $\mathbf{x}$ kan skrives

$$\mathbf{x} = \alpha_1\mathbf{v}_1 + \cdots + \alpha_m\mathbf{v}_m,$$

fås

$$R(\mathbf{x}) = \frac{\sum \alpha_i^2 \mathbf{v}_i'\mathbf{A}\,\mathbf{v}_i}{\sum \alpha_i^2 \mathbf{v}_i'\mathbf{B}\,\mathbf{v}_i} = \frac{\sum \lambda_i\alpha_i^2}{\sum \alpha_i^2}.$$

Heraf følger de to første påstande let. Hvis $\mathbf{x} \in M_k$, vil $\mathbf{x}$ have fremstillingen

$$\mathbf{x} = \alpha_k\mathbf{v}_k + \cdots + \alpha_m\mathbf{v}_m,$$

og

$$R(\mathbf{x}) = \frac{\lambda_k\alpha_k^2 + \cdots + \lambda_m\alpha_m^2}{\alpha_m^2 + \cdots + \alpha_m^2},$$

hvoraf resultatet følger. ■

1.4.5 Sporet af en matrix

Ved **sporet** af (den symmetriske) matrix $\mathbf{A}$ forstås summen af diagonalelementerne, i.e.

$$\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}.$$

For matricer $\mathbf{A}$ og $\mathbf{B}$ gælder (kvadr. matr.)

$$\mathrm{tr}(\mathbf{A}\,\mathbf{B}) = \mathrm{tr}(\mathbf{B}\,\mathbf{A}). \tag{1.5}$$

Endvidere haves, at sporet er lig summen af egenværdierne, i.e.

$$\mathrm{tr}(\mathbf{A}) = \sum_{i=1}^{n} \lambda_i.$$

Dette er en triviel følge af 1.5 og sætning 1.10
For positivt semidefinitte matricer er sporet derfor et andet mål for "størrelsen"af en matrix. Er sporet stort, er i det mindste nogle egenværdier store. Til gengæld er dette mål ikke følsomt overfor, om enkelte egenværdier er 0, d.v.s. om matricen er udartet. Det er determinanten derimod, idet vi jo erindrer, at

$$\det(\mathbf{A}) = \prod_{i=1}^{n} \lambda_i.$$

Vi bemærker ydermere, at for en idempotent matrix $\mathbf{A}$ gælder

$$\mathrm{tr}(\mathbf{A}) = \mathrm{rg}(\mathbf{A}).$$

Endvidere er

$$\mathrm{tr}(\mathbf{B}\,\mathbf{B}^{-}) = \mathrm{rg}(\mathbf{B}),$$

hvor $\mathbf{B}^{-}$ er en vilkårlig pseudoinvers til $\mathbf{B}$.

Til sidst nævner vi, at for en regulær matrix $\mathbf{S}$ gælder

$$\mathrm{tr}(\mathbf{S}^{-}\mathbf{B}\,\mathbf{S}) = \mathrm{tr}(\mathbf{B}).$$

1.4.6 Differentiation af linearform og kvadratisk form

Lad f : $R^n \to R$. Vi anvender da følgende skrivemåde for vektoren af partielle afledede

$$\frac{\partial \mathrm{f}(\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial \mathrm{f}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial \mathrm{f}(\mathbf{x})}{\partial x_1} \\ \vdots \\ \frac{\partial \mathrm{f}(\mathbf{x})}{\partial x_n} \end{bmatrix}.$$

Der gælder nu følgende sætning om differentiation af visse former

SÆTNING 1.24. For en symmetrisk $(n \times n)$-matrix $\mathbf{A}$ og en vilkårlig n-dimensional vektor $\mathbf{b}$ gælder

i) $\frac{\partial}{\partial \mathbf{x}}(\mathbf{b}'\mathbf{x}) = \mathbf{b}$

ii) $\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}'\mathbf{x}) = 2\mathbf{x}$

iii) $\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}'\mathbf{A}\,\mathbf{x}) = 2\mathbf{A}\,\mathbf{x}.$

▲

BEVIS 1.23. Beviset for i) og ii) er trivielt. iii) vises - bizart nok - bekvemmest ved hjælp af definitionen. For en vilkårlig vektor $\mathbf{h}$ haves

$$(\mathbf{x}+\mathbf{h})'\mathbf{A}(\mathbf{x}+\mathbf{h}) = \mathbf{x}'\mathbf{A}\,\mathbf{x} + \mathbf{h}'\mathbf{A}\,\mathbf{h} + 2\mathbf{h}'\mathbf{A}\,\mathbf{x}$$

Ved at vælge $\mathbf{h} = (0,\ldots,h,\ldots,0)'$ ses, at

$$\frac{\partial}{\partial \mathbf{x}_i}(\mathbf{x}'\mathbf{A}\,\mathbf{x}) = 2\sum_{j=1}^{h} a_{ij}x_j,$$

og heraf følger resultatet umiddelbart. ■

Vi belyser anvendelsen af sætningen i nedenstående

EKSEMPEL 1.8. Vi vil finde minimum af funktionen

$$\mathrm{g}(\theta) = (\mathbf{y} - \mathbf{A}\,\theta)'\mathbf{B}(\mathbf{y} - \mathbf{A}\,\theta),$$

hvor $\mathbf{y}$, $\mathbf{A}$ og $\mathbf{B}$ er givne og $\mathbf{B}$ endvidere positiv semidefinit (og symmetrisk). Da $\mathrm{g}(\theta)$ er konveks (en paraboloide, eventuelt udartet), kan minimumspunktet findes ved at løse ligningen

$$\frac{\partial}{\partial\theta}\mathrm{g}(\theta) = \mathbf{0}.$$

Vi foretager dog først en omskrivning af g. Vi har

$$\begin{aligned}\mathrm{g}(\theta) &= \mathbf{y}'\mathbf{B}\,\mathbf{y} - \theta'\mathbf{A}'\mathbf{B}\,\mathbf{y} + \theta'\mathbf{A}'\mathbf{B}\,\mathbf{A}\,\theta - \mathbf{y}'\mathbf{B}\,\mathbf{A}\,\theta \\ &= \mathbf{y}'\mathbf{B}\,\mathbf{y} - 2\mathbf{y}'\mathbf{B}\,\mathbf{A}\,\theta + \theta'\mathbf{A}'\mathbf{B}\,\mathbf{A}\,\theta.\end{aligned}$$

Her er benyttet, at

$$\theta'\mathbf{A}'\mathbf{B}\,\mathbf{y} = \mathbf{y}'\mathbf{B}\,\mathbf{A}\,\theta$$

(begge 1×1 matricer, d.v.s. en skalar, og hinandens transponerede). Heraf fås

$$\frac{\partial \mathrm{g}}{\partial\theta} = -2\mathbf{A}'\mathbf{B}\,\mathbf{y} + 2\mathbf{A}'\mathbf{B}\,\mathbf{A}\,\theta,$$

og det ses, at

$$\frac{\partial g}{\partial \theta} = \mathbf{0} \quad \leftrightarrow \quad \mathbf{A}'\mathbf{B}\,\mathbf{A}\theta = \mathbf{A}'\mathbf{B}\,\mathbf{y}.$$

Denne ligning har som sagt altid mindst en rod. Hvis $\mathbf{A}'\mathbf{B}\,\mathbf{A}$ er regulær fås

$$\theta_{\min} = (\mathbf{A}'\mathbf{B}\,\mathbf{A})^{-1}\mathbf{A}'\mathbf{B}\,\mathbf{y}.$$

Hvis matricen er singulær, kan vi skrive

$$\theta_{\min} = (\mathbf{A}'\mathbf{B}\,\mathbf{A})^{-}\mathbf{A}'\mathbf{B}\,\mathbf{y},$$

hvor $(\mathbf{A}'\mathbf{B}\,\mathbf{A})^{-}$ betegner en pseudoinvers til $\mathbf{A}'\mathbf{B}\,\mathbf{A}$. ♦

Vi kan nu finde en alternativ beskrivelse af hovedakserne i en ellisoide, idet vi har

SÆTNING 1.25. Lad $\mathbf{A}$ være en positiv definit symmetrisk matrix. Hovedretningerne i ellipsoiden E_c med ligningen

$$\mathbf{x}'\mathbf{A}\,\mathbf{x} = c, \qquad c > 0$$

er de retninger, hvor $\mathbf{x}'\mathbf{x}$, $\mathbf{x} \in E_c$, har stationære punkter. ▲

BEVIS 1.24. Vi kan antage, at $x = 1$. Vi skal da finde de stationære punkter for

$$\mathrm{f}(\mathbf{x}) = \mathbf{x}'\mathbf{x}$$

under bibetingelsen

$$\mathbf{x}'\mathbf{A}\,\mathbf{x} = 1$$

Vi anvender en Lagrange multiplikator teknik og definerer

$$\varphi(\mathbf{x}, \lambda) = \mathbf{x}'\mathbf{x} - \lambda(\mathbf{x}'\mathbf{A}\,\mathbf{x} - 1).$$

Ved differentiation fås

$$\frac{\partial \varphi}{\partial \mathbf{x}} = 2\mathbf{x} - 2\lambda\mathbf{A}\,\mathbf{x}.$$

Hvis denne størrelse skal være $\mathbf{0}$, må

$$\mathbf{x} = \lambda \mathbf{A}\,\mathbf{x}$$

eller

$$\mathbf{A}\,\mathbf{x} = \frac{1}{\lambda}\mathbf{x},$$

d.v.s. $\mathbf{x}$ må være en egenvektor. ■

1.5 Tensor- eller Kronecker produkt af matricer

Dette produkt kan med fordel anvendes ved en behandling af den flerdimensionale generelle lineære model.

DEFINITION 1.4. Lad $\mathbf{A}$ være en $m \times n$ matrix og $\mathbf{B}$ en $k \times \ell$ matrix. Ved **tensor** - eller **Kronecker-produktet** af $\mathbf{A}$ og $\mathbf{B}$ forstås matricen

$$\mathbf{A} \otimes \mathbf{B} = (a_{ij}\mathbf{B}) = \begin{bmatrix} a_{11}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ \vdots & & \vdots \\ a_{m1}\mathbf{B} & \cdots & a_{mn}\mathbf{B} \end{bmatrix} \tag{1.6}$$

Dette begreb svarer til tensorproduktet af lineære afbildninger, som kan gives en koordinatuafhængig fremstilling (se f.eks. [5]). Overføres dette til koordinatform kan enten anvendes 1.6 eller, hvad der er helt equivalent hermed, $\mathbf{A} \otimes \mathbf{B} = (\mathbf{A}b_{ij})$. Dette svarer blot til at ændre koordinaternes rækkefølge, d.v.s. til at ombytte rækker og søjler i de involverede matricer. ▲

Vi resumerer nu kort en række regneregler for tensorproduktet. Disse regneregler eftervises trivielt v.h.a. definitionen.

i) $\mathbf{O} \otimes \mathbf{A} = \mathbf{A} \otimes \mathbf{O} = \mathbf{O}$

ii) $(\mathbf{A}_1 + \mathbf{A}_2) \otimes \mathbf{B} = \mathbf{A}_1 \otimes \mathbf{B} + \mathbf{A}_2 \otimes \mathbf{B}$

iii) $\mathbf{A} \otimes (\mathbf{B}_1 + \mathbf{B}_2) = \mathbf{A} \otimes \mathbf{B}_1 + \mathbf{A} \otimes \mathbf{B}_2$

iv) $\alpha\mathbf{A} \otimes \beta\mathbf{B} = \alpha\beta\mathbf{A} \otimes \mathbf{B}$

v) $\mathbf{A}_1\mathbf{A}_2 \otimes \mathbf{B}_1\mathbf{B}_2 = (\mathbf{A}_1 \otimes \mathbf{B}_1)(\mathbf{A}_2 \otimes \mathbf{B}_2)$

vi) $(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$, hvis de inverse eksisterer

vii) $(\mathbf{A} \otimes \mathbf{B})^{-} = \mathbf{A}^{-} \otimes \mathbf{B}^{-}$

viii) $(\mathbf{A} \otimes \mathbf{B})' = \mathbf{A}' \otimes \mathbf{B}'$

ix) Lad $\mathbf{A}$, symmetrisk og $p \times p$, have egenværdier $\alpha_1, \ldots, \alpha_p$ og egenvektorer $\mathbf{x}_i$, og lad $\mathbf{B}$, symmetrisk og $q \times q$, have egenværdier $\beta_1, \ldots, \beta_q$ og egenvektorer $\mathbf{y}_1, \ldots, \mathbf{y}_q$. Da vil $\mathbf{A} \otimes \mathbf{B}$ have egenværdierne $\alpha_i\beta_j$, $i = 1, \ldots, p$, $j = 1, \ldots, q$, med tilsvarende egenvektorer.

$$(\mathbf{x}_i \otimes \mathbf{y}_j \sim) \begin{bmatrix} x_{1i}\mathbf{y}_j \\ \vdots \\ x_{pi}\mathbf{y}_j \end{bmatrix}$$

x) $\det(\mathbf{A} \otimes \mathbf{B}) = (\det \mathbf{A})^q (\det \mathbf{B})^p$

1.6 Indre produkter og normer

For n-dimensionale vektorer har man som bekendt defineret det **indre produkt** eller **skalarproduktet** eller **prikproduktet** af $\mathbf{x}$ og $\mathbf{y}$ ved

$$\mathbf{x} \cdot \mathbf{y} = \mathbf{x}'\mathbf{y} = (x_1 \ldots x_n) \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \sum_{i=n}^{n} x_i y_i,$$

og man bemærker, at $\mathbf{x}$ og $\mathbf{y}$ er **ortogonale**, hvis og kun hvis

$$\mathbf{x} \cdot \mathbf{y} = \mathbf{x}'\mathbf{y} = 0.$$

Den tilhørende norm er

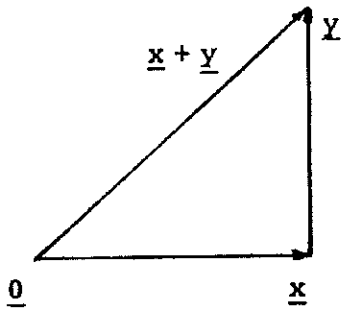
$$\|\mathbf{x}\| = (\mathbf{x} \cdot \mathbf{x})^{\frac{1}{2}} = (\mathbf{x}'\mathbf{x})^{\frac{1}{2}} = \sqrt{x_1^2 + \cdots + x_n^2}$$

Vi bemærker, at $\|\mathbf{x} - \mathbf{y}\|$ angiver den **euklidiske afstand** mellem punkterne $\mathbf{x}$ og $\mathbf{y}$.

For **ortogonale** vektorer $\mathbf{x}$ og $\mathbf{y}$ (d.v.s. $\mathbf{x} \perp \mathbf{y}$) har vi den pythagoræiske læresætning

$$\|\mathbf{x} + \mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2;$$

se skitsen. Endvidere bemærker vi, at projektionen $p(\mathbf{x})$ (den ortogonale) af en vektor

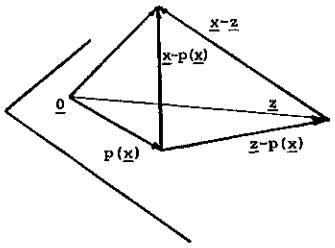


$$\begin{aligned}
&\|\mathbf{x}+\mathbf{y}\|^2 \\
&= (\mathbf{x}+\mathbf{y})'(\mathbf{x}+\mathbf{y}) \\
&= \mathbf{x}'\mathbf{x} + \mathbf{x}'\mathbf{y} + \mathbf{y}'\mathbf{x} + \mathbf{y}'\mathbf{y} \\
&= \mathbf{x}'\mathbf{x} + \mathbf{y}'\mathbf{y} \\
&= \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2.
\end{aligned}$$

$\mathbf{x}$ ned på et underrum U kan bestemmes ved hjælp af normen, idet vi har, at $p(\mathbf{x})$ er givet ved

$$\|\mathbf{x} - p(\mathbf{x})\| = \min_{\mathbf{z}\in U} \|\mathbf{x} - \mathbf{z}\|$$

BEVIS 1.25.



Ifølge den pythagoræiske læresætning har vi, at

$$\begin{aligned}
&\|\mathbf{x} - p(\mathbf{x})\|^2 - \|\mathbf{z} - p(\mathbf{x})\|^2 \\
&= \|\mathbf{x} - \mathbf{z}\|^2,
\end{aligned}$$

d.v.s. minimumsværdien af

$$= \|\mathbf{x} - \mathbf{z}\|^2,$$ og dermed af

$$= \|\mathbf{x} - \mathbf{z}\|$$ indtræffer for

$\mathbf{z} = p(\mathbf{x})$. ■

Det er nu meget let at vise, at gyldigheden af ovenstående resultater blot beror på 4 grundlæggende egenskaber ved det indre produkt, nemlig, idet vi nu betegner det indre produkt af $\mathbf{x}$ og $\mathbf{y}$ med $(\mathbf{x}|\mathbf{y})$.

$$\begin{aligned}
&\text{IP1}: \quad (\mathbf{x}|\mathbf{y}) = (\mathbf{y}|\mathbf{x}) \\
&\text{IP2}: \quad (\mathbf{x}+\mathbf{y}|\mathbf{z}) = (\mathbf{x}|\mathbf{z}) + (\mathbf{y}|\mathbf{z}) \\
&\text{IP3}: \quad (k\mathbf{x}|\mathbf{y}) = k(\mathbf{x}|\mathbf{y}) \\
&\text{IP4}: \quad \mathbf{x} \neq \mathbf{0} \Rightarrow (\mathbf{x}|\mathbf{x}) > 0.
\end{aligned}$$

For en vilkårlig bilinearform $(\cdot|\cdot)$, der tilfredsstiller ovenstående, kan man definere et ortogonalitetsbegreb ved

$$\mathbf{x} \perp \mathbf{y} \quad \overset{d}{\Leftrightarrow} \quad (\mathbf{x}|\mathbf{y}) = 0.$$

For en vilkårlig positiv definit symmetrisk matrix $\mathbf{A}$ kan vi definere et indre produkt ved

$$(\mathbf{x}|\mathbf{y})_{\mathbf{A}} = \mathbf{x}'\mathbf{A}\,\mathbf{y}.$$

Det er helt trivielt at eftervise, at IP 1-4 er opfyldte. For dette indre produkt og den deraf fødte norm givet ved

$$\|\mathbf{x}\|_{\mathbf{A}} = \sqrt{(\mathbf{x}|\mathbf{x})_{\mathbf{A}}} = \sqrt{\mathbf{x}'\mathbf{A}\,\mathbf{x}},$$

vil vi, hvor misforståelse er udelukket, anvende betegnelserne $(\mathbf{x}|\mathbf{y})$ og $\|\mathbf{x}\|$.

Man bemærker, at mængden af punkter med konstant $\mathbf{A}$-norm lig 1 er mængden

$$\{\mathbf{x}|\ \|\mathbf{x}\|^2 = 1\} = \{\mathbf{x}|\mathbf{x}'\mathbf{A}\,\mathbf{x} = 1\},$$

d.v.s. punkterne på en ellipsoide.

Omvendt svarer der til enhver ikke udartet ellipsoide en symmetrisk positivt definit matrix $\mathbf{A}$, så

$$E = \{\mathbf{x}|\mathbf{x}'\mathbf{A}\,\mathbf{x} = 1\} = \{\mathbf{x}|\ \|\mathbf{x}\|_{\mathbf{A}}^2 = 1\}.$$

Der er på denne måde tilvejebragt en forbindelse mellem mængden af mulige indre produkter og mængden af ellipsoider.

To vektorer $\mathbf{x}$ og $\mathbf{y}$ er **ortogonale (med hensyn til $\mathbf{A}$)**, hvis

$$\mathbf{x}'\mathbf{A}\,\mathbf{y} = 0,$$

d.v.s. hvis $\mathbf{x}$ og $\mathbf{y}$ er **konjugerede retninger** i ellipsoiden svarende til $\mathbf{A}$.

Man kan også indføre et **vinkelbegreb** v.h.a. definitionen

$$\cos(\angle\mathbf{a},\mathbf{b}) = \frac{(\mathbf{a}|\mathbf{b})}{\|\mathbf{a}\|\,\|\mathbf{b}\|}.$$

Vi anfører nu et lemma, som skal anvendes i forbindelse med sætninger om uafhængighed af projektioner af normalt fordelte stokastiske variable.

LEMMA 1.1. Lad R^n være spaltet i en direkte sum

$$R^n = U_1 \oplus \cdots \oplus U_k$$

af n_i dimensionale underrum, der er ortogonale m.h.t. den positivt definitte matrix Σ^{-1}, d.v.s.

$$\mathbf{x} \perp \mathbf{y} \Leftrightarrow \mathbf{x}'\Sigma^{-1}\mathbf{y} = 0.$$

Lad for $i = 1, \ldots, k$ projektionen p_i på U_i være givet ved matricen $\mathbf{C}_i$. Da vil

$$\mathbf{C}_i \Sigma \mathbf{C}_j' = 0$$

for alle $i \neq j$. Endvidere gælder

$$\Sigma^{-1}\mathbf{C}_i = \mathbf{C}_i'\Sigma^{-1} = \mathbf{C}_i'\Sigma\,\mathbf{C}_i.$$

BEVIS 1.26. Da $p_i \circ p_i = p_i$, gælder

$$\mathbf{C}_i\mathbf{C}_i = \mathbf{C}_i,$$

og da

$$p_i(\mathbf{x}) \perp \mathbf{x} - p_i(\mathbf{x}),$$

(se skitsen) vil

$$p_i(\mathbf{x})'\Sigma^{-1}(\mathbf{x} - p_i(\mathbf{x})) = 0,$$

d.v.s.

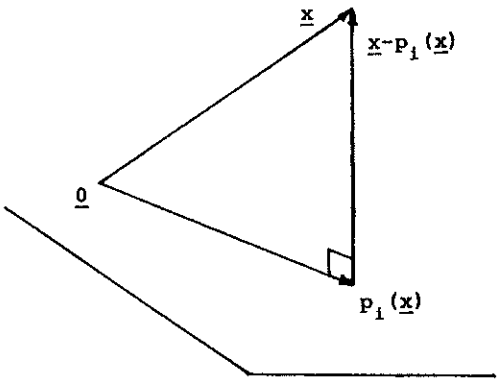
$$\mathbf{x}\mathbf{C}_i'\Sigma^{-1}[\mathbf{x} - \mathbf{C}_i\mathbf{x}] = 0.$$

Dette gælder for alle $\mathbf{x}$, og derfor er

$$\mathbf{C}_i'\Sigma^{-1}(\mathbf{I} - \mathbf{C}_i) = \mathbf{0},$$

eller

$$\mathbf{C}_i'\Sigma^{-1} = \mathbf{C}_i'\Sigma^{-1}\mathbf{C}_i.$$



Nu er højresiden klart symmetrisk, hvorfor vi får

$$\mathbf{C}_i'\Sigma^{-1} = \Sigma^{-1}\mathbf{C}_i.$$

Ved at multiplicere foran og bagved med Σ fås

$$\Sigma\,\mathbf{C}_i' = \mathbf{C}_i\Sigma,$$

hvorfor

$$\mathbf{C}_i\boldsymbol{\Sigma}\,\mathbf{C}_i' = \mathbf{C}_i\mathbf{C}_i\boldsymbol{\Sigma} = \mathbf{C}_i\boldsymbol{\Sigma}.$$

Dette giver

$$\mathbf{C}_i\boldsymbol{\Sigma}\,\mathbf{C}_j' = \mathbf{C}_i\boldsymbol{\Sigma}\,\mathbf{C}_i'\mathbf{C}_j' = \mathbf{C}_i\boldsymbol{\Sigma}\,\mathbf{0} = \mathbf{0}.$$

Det næstsidste lighedstegn følger af, at summen er direkte, hvorfor det for ethvert $\mathbf{x}$ gælder

$$p_j(p_i(\mathbf{x})) = \mathbf{0},$$

d.v.s.

$$\mathbf{C}_j\mathbf{C}_i\mathbf{x} = \mathbf{0}.$$

Da $\mathbf{x}$ som nævnt er vilkårlig, medfører dette

$$\mathbf{C}_j\mathbf{C}_i = \mathbf{0},$$

eller

$$\mathbf{C}_i'\mathbf{C}_j' = \mathbf{0}.$$

■

Kapitel 2
Flerdimensionale variable

I dette kapitel gives indledningsvis en supplering af de resultater om flerdimensionale stokastiske variable, der er anført i kapitel 0 i bind 1. Dernæst omtales specielt den flerdimensionale, normale fordeling og fordelinger afledt af denne. Endelig gives en kort beskrivelse af de særlige forhold, som estimation og testning af flere parametre simultant giver anledning til.

2.1 Momenter af flerdimensionale stokastiske variable

Vi indleder med

2.1.1 Middelværdi

Lad der være givet en **stokastisk matrix** i.e. en matrix, hvis enkelte elementer er stokastiske variable:

$$\mathbf{X} = \begin{bmatrix} X_{11} & \cdots & X_{1n} \\ \vdots & & \vdots \\ X_{k1} & \cdots & X_{kn} \end{bmatrix}$$

Vi definerer da **middelværdien** eller **forventningsværdien** af $\mathbf{X}$ ved

$$\mathrm{E}(\mathbf{X}) = \begin{bmatrix} \mathrm{E}(X_{11}) & \cdots & \mathrm{E}(X_{1n}) \\ \vdots & & \vdots \\ \mathrm{E}(X_{k1}) & \cdots & \mathrm{E}(X_{kn}) \end{bmatrix} = \begin{bmatrix} \mu_{11} & \cdots & \mu_{1n} \\ \vdots & & \vdots \\ \mu_{k1} & \cdots & \mu_{kn} \end{bmatrix} = \mu.$$

SÆTNING 2.1. Lad $\mathbf{A}$ være en $k \times n$ matrix af konstanter. Da er

$$\mathrm{E}(\mathbf{A} + \mathbf{X}) = \mathbf{A} + \mathrm{E}(\mathbf{X}).$$

Denne sætning er lige som den følgende en triviel følge af definitionen. ▲

SÆTNING 2.2. Lad $\mathbf{A}$ og $\mathbf{B}$ være konstante stokastiske matricer, således at $\mathbf{A}\,\mathbf{x}$ og $\mathbf{x}\,\mathbf{B}$ eksisterer. Da er

$$\begin{aligned} \mathrm{E}(\mathbf{A}\,\mathbf{X}) &= \mathbf{A}\,\mathrm{E}(\mathbf{X}) \\ \mathrm{E}(\mathbf{X}\,\mathbf{B}) &= \mathrm{E}(\mathbf{X})\mathbf{B} \end{aligned}$$

▲

Endelig har vi

SÆTNING 2.3. Lad $\mathbf{X}$ og $\mathbf{Y}$ være stokastiske matricer af samme orden. Da gælder

$$\mathrm{E}(\mathbf{X} + \mathbf{Y}) = \mathrm{E}(\mathbf{X}) + \mathrm{E}(\mathbf{Y}).$$

▲

BEMÆRKNING 2.1. Vi har intetsteds nævnt, at vi naturligvis forudsætter, at de involverede forventningsværdier eksisterer. Dette forudsættes her og overalt i det følgende, hvor disse begreber måtte blive nævnt. ▼

2.1.2 Dispersionsmatricen

Generaliseringen af variansen af en stokastisk variabel er begrebet **dispersionsmatricen** for en stokastisk vektor $\mathbf{X}$. Den defineres ved

$$\mathrm{D}(\mathbf{X}) = \boldsymbol{\Sigma} = \mathrm{E}\{(\mathbf{X} - \mu)\,(\mathbf{X} - \mu)'\},$$

hvor

$$\mu = \mathrm{E}(\mathbf{X}).$$

Det skal bemærkes, at $\mathrm{D}(\mathbf{X})$ også ofte omtales som **kovariansmatricen** og da betegnes $\mathrm{Cov}(\mathbf{X})$. Dette er dog en lidt uheldig vending, idet den vil kunne give anledning til forveksling med kovariansen mellem to (flerdimensionale) stokastiske variable. En anden hyppigt anvendt betegnelse er $\mathrm{V}(\mathbf{X})$. Vi bemærker i øvrigt, at

$$(\mathbf{X}-\mu)(\mathbf{X}-\mu)' = \begin{bmatrix} X_1-\mu_1 \\ \vdots \\ X_n-\mu_n \end{bmatrix} (X_1-\mu_1,\ldots,X_n-\mu_n) =$$

$$\begin{bmatrix} (X_1-\mu_1)^2 & (X_1-\mu_1)(X_2-\mu_2) & \cdots & (X_1-\mu_1)(X_n-\mu_n) \\ (X_2-\mu_2)(X_1-\mu_1) & (X_2-\mu_2)^2 & \cdots & (X_2-\mu_2)(X_n-\mu_n) \\ \vdots & \vdots & \ddots & \vdots \\ (X_n-\mu_n)(X_1-\mu_1) & (X_n-\mu_n)(X_2-\mu_2) & \cdots & (X_n-\mu_n)^2 \end{bmatrix}$$

d.v.s. dispersionsmatricens (i,j)'te element er $\mathrm{Cov}(X_i,X_j)$, eller

$$\boldsymbol{\Sigma} = \mathrm{D}(\mathbf{X}) = \begin{bmatrix} \mathrm{V}(X_1) & \mathrm{Cov}(X_1,X_2) & \cdots & \mathrm{Cov}(X_1,X_n) \\ \mathrm{Cov}(X_2,X_1) & \mathrm{V}(X_2) & \cdots & \mathrm{Cov}(X_2,X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n,X_1) & \mathrm{Cov}(X_n,X_2) & \cdots & \mathrm{V}(X_n) \end{bmatrix}.$$

Vi vil ofte anvende skrivemåden

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{bmatrix} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2n} \\ \vdots & \vdots & & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_{nn} \end{bmatrix},$$

idet vi om varianserne såvel anvender betegnelserne σ_i^2 som σ_{ii}. Vi har, at $\boldsymbol{\Sigma}$ er **symmetrisk**. Mere interessant er

SÆTNING 2.4. Dispersionsmatricen $\boldsymbol{\Sigma}$ for en stokastisk vektor (d.v.s. flerdimensional stokastisk variabel) er positivt semidefinit. ▲

BEVIS 2.1. For en vilkårlig vektor $\mathbf{y}$ haves

$$\begin{aligned}\mathbf{y}'\,\Sigma\,\mathbf{y} &= \mathbf{y}'\,\mathrm{E}\{(\mathbf{X}-\mu)(\mathbf{X}-\mu)'\}\mathbf{y}\\ &= \mathrm{E}\{\mathbf{y}'\,(\mathbf{X}-\mu)(\mathbf{X}-\mu)'\mathbf{y}\}\\ &= \mathrm{E}\{\,[(\mathbf{X}-\mu)'\mathbf{y}]'[(\mathbf{X}-\mu)'\mathbf{y}]\,\}\\ &\geq 0\,,\end{aligned}$$

da størrelsen i de krøllede paranteser er ≥ 0. ■

Der gælder sætninger ganske analoge til de fra de endimensionale stokastiske variable kendte.

SÆTNING 2.5. Lad $\mathbf{X}$ og $\mathbf{Y}$ være uafhængige. Da er

$$\mathrm{D}(\mathbf{X}+\mathbf{Y}) = \mathrm{D}(\mathbf{X}) + \mathrm{D}(\mathbf{Y}).$$

Lad $\mathbf{b}$ være en konstant. Da har vi

$$\mathrm{D}(\mathbf{b}+\mathbf{X}) = \mathrm{D}(\mathbf{X}).$$

Hvis $\mathbf{A}$ er en konstant matrix, således at $\mathbf{A}\,\mathbf{X}$ eksisterer, da gælder

$$\mathrm{D}(\mathbf{A}\,\mathbf{X}) = \mathbf{A}\;\mathrm{D}(\mathbf{X})\mathbf{A}'.$$

▲

BEVIS 2.2. Den første relation følger af

$$\begin{aligned}\mathrm{Cov}(X_i+Y_i, X_j+Y_j) &= \mathrm{Cov}(X_i,X_j) + \mathrm{Cov}(X_i,Y_j) +\\ &\quad\ \mathrm{Cov}(Y_i,X_j) + \mathrm{Cov}(Y_i,Y_j)\\ &= \mathrm{Cov}(X_i,X_j) + \mathrm{Cov}(Y_i,Y_j),\end{aligned}$$

idet $\mathrm{Cov}(Y_i, X_j) = 0$, da X_j og Y_i er uafhængige. Den næste relation er triviel. Den sidste følger af

$$\begin{aligned}\mathrm{D}(\mathbf{A}\,\mathbf{X}) &= \mathrm{E}\{(\mathbf{A}\,\mathbf{X}-\mathbf{A}\,\mu)(\mathbf{A}\,\mathbf{X}-\mathbf{A}\,\mu)'\}\\ &= \mathrm{E}\{\mathbf{A}[\mathbf{X}-\mu][\mathbf{X}-\mu]'\mathbf{A}'\}\\ &= \mathbf{A}\,\mathrm{E}\{[\mathbf{X}-\mu][\mathbf{X}-\mu]'\}\mathbf{A}'\\ &= \mathbf{A}\,\mathrm{D}(\mathbf{X})\mathbf{A}'\\ &= \mathbf{A}\,\Sigma\,\mathbf{A}'\end{aligned}$$

■

Sætter vi

$$\mathbf{V} = \operatorname{diag}\left(\frac{1}{\sigma_1}, \ldots, \frac{1}{\sigma_n}\right) = \begin{bmatrix} \sigma_1^{-1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_n^{-1} \end{bmatrix}$$

og "skalerer" vi $\mathbf{X}$ med $\mathbf{V}$, fås

$$\mathrm{D}(\mathbf{V}\,\mathbf{X}) = \mathbf{V}\,\boldsymbol{\Sigma}\,\mathbf{V}' = \begin{bmatrix} 1 & \frac{\sigma_{12}}{\sigma_1\,\sigma_2} & \cdots & \frac{\sigma_{1n}}{\sigma_1\,\sigma_n} \\ \frac{\sigma_{12}}{\sigma_1\,\sigma_2} & 1 & \cdots & \frac{\sigma_{2n}}{\sigma_2\,\sigma_n} \\ \vdots & \vdots & & \vdots \\ \frac{\sigma_{1n}}{\sigma_1\,\sigma_n} & \frac{\sigma_{2n}}{\sigma_2\,\sigma_n} & \cdots & 1 \end{bmatrix}.$$

Vi ser, at elementerne netop er korrelationskoefficienterne mellem $\mathbf{X}_j'$ komponenter, hvorfor denne matrix også benævnes **korrelationsmatricen** for $\mathbf{X}$, og vi skriver

$$R(\mathbf{X}) = \begin{bmatrix} 1 & \cdots & \rho_{1n} \\ \vdots & & \vdots \\ \rho_{1n} & \cdots & 1 \end{bmatrix},$$

hvor altså

$$\rho_{ij} = \operatorname{Cor}(X_i, X_j) = \frac{\operatorname{Cov}(X_i, X_j)}{\sqrt{\mathrm{V}(X_i)\,\mathrm{V}(X_j)}}.$$

2.1.3 Kovarians

Lad der være givet to stokastiske variable

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix} \quad \text{og} \quad \mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_q \end{bmatrix}$$

med middelværdier μ og ν. Vi definerer da **kovariansen** mellem $\mathbf{X}$ og $\mathbf{Y}$ ved

$$\mathrm{C}(\mathbf{X}, \mathbf{Y}) = \mathrm{E}[(\mathbf{X} - \mu)(\mathbf{Y} - \nu)'] = \begin{bmatrix} \operatorname{Cov}(X_1, Y_1) & \cdots & \operatorname{Cov}(X_1, Y_q) \\ \vdots & & \vdots \\ \operatorname{Cov}(X_p, Y_1) & \cdots & \operatorname{Cov}(X_p, Y_q) \end{bmatrix}.$$

Der gælder da

$$\mathrm{C}(\mathbf{X},\mathbf{X}) = \mathrm{D}(\mathbf{X})$$

og

$$\mathrm{C}(\mathbf{X},\mathbf{Y}) = [\mathrm{C}(\mathbf{Y},\mathbf{X})]'.$$

Mere dyb er

SÆTNING 2.6. Lad $\mathbf{X}$ og $\mathbf{Y}$ være som ovenfor, og lad $\mathbf{A}$ og $\mathbf{B}$ være henholdsvis $n \times p$ og $m \times q$ matricer af konstanter. Da er

$$\mathrm{C}(\mathbf{A}\,\mathbf{X},\mathbf{B}\,\mathbf{Y}) = \mathbf{A}\,\mathrm{C}(\mathbf{X},\mathbf{Y})\mathbf{B}'.$$

Er $\mathbf{U}$ en p-dimensional og $\mathbf{V}$ en q-dimensional stokastisk variabel gælder endvidere

$$\mathrm{C}(\mathbf{X}+\mathbf{U},\mathbf{Y}) = \mathrm{C}(\mathbf{X},\mathbf{Y}) + \mathrm{C}(\mathbf{U},\mathbf{Y})$$

$$\mathrm{C}(\mathbf{X},\mathbf{Y}+\mathbf{V}) = \mathrm{C}(\mathbf{X},\mathbf{Y}) + \mathrm{C}(\mathbf{X},\mathbf{V}).$$

Endelig gælder,

$$\mathrm{D}(\mathbf{X}+\mathbf{U}) = \mathrm{D}(\mathbf{X}) + \mathrm{D}(\mathbf{U}) + \mathrm{C}(\mathbf{X},\mathbf{U}) + \mathrm{C}(\mathbf{U},\mathbf{X}).$$

▲

BEVIS 2.3. Vi har ifølge definitionen

$$\begin{aligned}
\mathrm{C}(\mathbf{A}\,\mathbf{X},\mathbf{B}\,\mathbf{Y}) &= \mathrm{E}[(\mathbf{A}\,\mathbf{X}-\mathbf{A}\,\mu)(\mathbf{B}\,\mathbf{Y}-\mathbf{B}\,\nu)'] \\
&= \mathrm{E}[\mathbf{A}(\mathbf{X}-\mu)(\mathbf{Y}-\nu)'\mathbf{B}'] \\
&= \mathbf{A}\,\mathrm{E}[(\mathbf{X}-\mu)(\mathbf{Y}-\nu)']\mathbf{B}' \\
&= \mathbf{A}\,\mathrm{C}(\mathbf{X},\mathbf{Y})\mathbf{B}'.
\end{aligned}$$

Hermed er den første påstand godtgjort. Tilsvarende fås - idet vi sætter $E(U) = \delta$ -

$$\begin{aligned}
\mathrm{C}(\mathbf{X}+\mathbf{U},\mathbf{Y}) &= \mathrm{E}[(\mathbf{X}+\mathbf{U}-\mu-\delta)(\mathbf{Y}-\nu)'] \\
&= \mathrm{E}[(\mathbf{X}-\mu)(\mathbf{Y}-\nu)' + (\mathbf{U}-\delta)(\mathbf{Y}-\nu)'] \\
&= \mathrm{E}[(\mathbf{X}-\mu)(\mathbf{Y}-\nu)'] + \mathrm{E}[(\mathbf{U}-\delta)(\mathbf{Y}-\nu)'] \\
&= \mathrm{C}(\mathbf{X},\mathbf{Y}) + \mathrm{C}(\mathbf{U},\mathbf{Y}),
\end{aligned}$$

og den tilsvarende relation med $\mathbf{Y} + \mathbf{V}$ vises helt analogt. Endelig har vi

$$\begin{aligned} \mathrm{D}(\mathbf{X}+\mathbf{U}) &= \mathrm{C}(\mathbf{X}+\mathbf{U}, \mathbf{X}+\mathbf{U}) \\ &= \mathrm{C}(\mathbf{X},\mathbf{X}) + \mathrm{C}(\mathbf{X},\mathbf{U}) + \mathrm{C}(\mathbf{U},\mathbf{X}) + \mathrm{C}(\mathbf{U},\mathbf{U}). \end{aligned}$$

■

Hvis $C(\mathbf{X}, \mathbf{Y}) = \mathbf{0}$ siges $\mathbf{X}$ og $\mathbf{Y}$ at være **ukorrelerede**. Det svarer til at alle komponenter i $\mathbf{X}$ er ukorrelerede med samtlige komponenter i $\mathbf{Y}$.

Ved behandlingen af den flerdimensionale generelle lineære model får vi brug for følgende

SÆTNING 2.7. Lad $\mathbf{X}_1, \ldots, \mathbf{X}_n$ være uafhængige, p-dimensionale stokastiske variable med samme dispersionsmatrix $\boldsymbol{\Sigma} = (\sigma_{ij})$. Vi sætter

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1' \\ \vdots \\ \mathbf{X}_n' \end{bmatrix} = \begin{bmatrix} X_{11} & \cdots & X_{p1} \\ \vdots & & \vdots \\ X_{1n} & \cdots & X_{pn} \end{bmatrix}$$

(Bemærk, at variabel-index er det første index og gentagelsesnr.-index er det andet). Defineres

$$\mathrm{vc}(\mathbf{X}) = \begin{bmatrix} X_{11} \\ \vdots \\ X_{1n} \\ \vdots \\ X_{p1} \\ \vdots \\ X_{pn} \end{bmatrix}$$

d.v.s. som vektoren bestående af søjlerne i $\mathbf{X}$ (vc = vector og columns) fås

$$\mathrm{D}(\mathrm{vc}(\mathbf{X})) = \boldsymbol{\Sigma} \otimes \mathbf{I}_n,$$

hvor $\mathbf{I}_n$ er enhedsmatricen af n'te orden. ▲

BEVIS 2.4. Følger trivielt af definitionen på tensorprodukt og af definitionen på dispersionsmatricen. ■

2.2 Den flerdimensionale normalfordeling

Den flerdimensionale normalfordeling spiller samme vigtige rolle i teorien for flerdimensionale variable, som normalfordelingen gør i det endimensionale tilfælde. Vi indleder med

2.2.1 Definition og simple egenskaber

Lad $X_1, \ldots, X_p$ være indbyrdes uafhængige, $\mathrm{N}(0,1)$ - fordelte variable. Vi siger da, at

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_p \end{bmatrix},$$

er **standardiseret (normeret) p-dimensionalt normalt fordelt**, og vi skriver

$$\mathbf{X} \in \mathrm{N}(\mathbf{0}, \mathbf{I}) = \mathrm{N}_p(\mathbf{0}, \mathbf{I}),$$

hvor den sidste betegnelse anvendes, hvis der kan opstå tvivl om dimensionen. Vi bemærker, at

$$\mathrm{E}(\mathbf{X}) = \mathbf{0}, \quad \mathrm{D}(\mathbf{X}) = \mathbf{I}.$$

Vi definerer den flerdimensionale normalfordeling med generelle parametre i

DEFINITION 2.1. Vi siger, at den p-dimensionale stokastiske variabel $\mathbf{X}$ er **normalt fordelt med parametre** μ og Σ, hvis $\mathbf{X}$ har samme fordeling som

$$\mu + \mathbf{A}\,\mathbf{U},$$

hvor $\mathbf{A}$ tilfredsstiller

$$\mathbf{A}\,\mathbf{A}' = \Sigma,$$

og hvor $\mathbf{U}$ er standardiseret p-dimensionalt normalt fordelt. Vi skriver

$$\mathbf{X} \in \mathrm{N}(\mu, \Sigma) = \mathrm{N}_p(\mu, \Sigma),$$

hvor den sidste betegnelse igen anvendes, hvis der kan være tvivl om dimensionen. ▲

BEMÆRKNING 2.2. Definitionen er kun lovlig, hvis man viser, at $\mathbf{A}\,\mathbf{A}' = \mathbf{B}\,\mathbf{B}'$ medfører

$$\mathfrak{L}(\mu + \mathbf{A}\,\mathbf{U}) = \mathfrak{L}(\mu + \mathbf{B}\,\mathbf{V}),$$

hvor $\mathbf{U}$ og $\mathbf{V}$ er standardiseret normalt fordelte og ikke nødvendigvis af samme dimension. Relationen er gyldig, men dette vil vi dog ikke komme nærmere ind på her. Af sætning 1.10 fås umiddelbart, at der for en vilkårlig positivt semidefinit matrix Σ eksisterer en matrix $\mathbf{A}$ med $\mathbf{A}\,\mathbf{A}' = \Sigma$, således at betegnelsen $\mathrm{N}(\mu, \Sigma)$ har mening for en vilkårlig positivt semidefinit $p \times p$ matrix Σ og en vilkårlig p-dimensional vektor μ.

Man bemærker trivielt at

$$\mathbf{X} \in \mathrm{N}(\mu, \Sigma) \quad \Rightarrow \quad \text{i) } \mathrm{E}(\mathbf{X}) = \mu \quad \wedge \quad \text{ii) } \mathrm{D}(\mathbf{X}) = \Sigma$$

d.v.s. fordelingen er parametriseret ved dens middelværdi og dispersionsmatrix. ▼

Hvis Σ har fuld rang har fordelingen en tæthed, som vi anfører i

SÆTNING 2.8. Lad $\mathbf{X} \in \mathrm{N}_p(\mu, \Sigma)$, og lad $\mathrm{rg}(\Sigma) = p$. Da har $\mathbf{X}$ tætheden

$$\begin{aligned} \mathrm{f}(\mathbf{x}) &= \frac{1}{\sqrt{2\pi}^p} \frac{1}{\sqrt{\det \Sigma}} \exp[-\frac{1}{2}(\mathbf{x}-\mu)'\Sigma^{-1}(\mathbf{x}-\mu)] \\ &= \frac{1}{\sqrt{2\pi}^p} \frac{1}{\sqrt{\det \Sigma}} \exp[-\frac{1}{2}\|\mathbf{x}-\mu\|_1^2], \end{aligned}$$

hvor den anførte norm er den ved Σ^{-1} bestemte, jvf. p. 53. ▲

BEVIS 2.5. Lad $\mathbf{U} \in \mathrm{N}_p(\mathbf{0}, \mathbf{I})$. Da har $\mathbf{U}$ tætheden

$$\begin{aligned} \mathrm{h}(\mathbf{u}) &= \prod_{i=1}^{p} \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}u_i^2) = \frac{1}{\sqrt{2\pi}^p} \exp(-\frac{1}{2}\sum_{i=1}^{p} u_i^2) \\ &= \frac{1}{\sqrt{2\pi}^p} \exp(-\frac{1}{2}\mathbf{u}'\,\mathbf{u}). \end{aligned}$$

Vi betragter dernæst transformationen fra $R^p \to R^p$ givet ved

$$\mathbf{u} \to \mathbf{x} = \mu + \mathbf{A}\,\mathbf{u}$$

hvor $\mathbf{A}\,\mathbf{A}' = \boldsymbol{\Sigma}$. Af sætning 1.14 følger, at $\mathbf{A}$ er regulær. Vi får

$$\mathbf{u} = \mathbf{A}^{-1}(\mathbf{x} - \mu),$$

hvorfor

$$\begin{aligned} \mathbf{u}'\mathbf{u} &= (\mathbf{x} - \mu)'\mathbf{A}^{-1\prime}\mathbf{A}^{-1}(\mathbf{x} - \mu) \\ &= (\mathbf{x} - \mu)'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mu). \end{aligned}$$

Da endvidere

$$\det(\boldsymbol{\Sigma}) = \det(\mathbf{A}\,\mathbf{A}') = \det(\mathbf{A})^2,$$

d.v.s

$$\det(\mathbf{A}^{-1}) = \frac{1}{\sqrt{\det \boldsymbol{\Sigma}}}$$

følger resultatet af sætning 0.8 i bind 1. ■

Vi bemærker at den inverse dispersionsmatrix $\boldsymbol{\Sigma}^{-1}$ ofte kaldes **præcisionen** for den normale fordeling.

Hvis $\boldsymbol{\Sigma}$ ikke er regulær, er fordelingen udartet og har ingen tæthed. Vi indfører da begrebet den affine støtte i

DEFINITION 2.2. Lad $\mathbf{X} \in \mathrm{N}_p(\mu, \boldsymbol{\Sigma})$. Ved den (**affine**) **støtte** for $\mathbf{X}$ forstås det mindste (side-) underrum af R^p, hvor $\mathbf{X}$ er defineret med sandsynlighed 1. ▲

BEMÆRKNING 2.3. Hvis man indskrænker betragtningerne til den affine støtte, er $\mathbf{X}$ regulært fordelt og har da en tæthed som angivet i sætning 2.8. ▼

Vi har forskellige muligheder for at bestemme støtten for en p-dimensional normalfordeling. Der er først

SÆTNING 2.9. Lad $\mathbf{X} \in \mathrm{N}_p(\mu, \boldsymbol{\Sigma})$, og lad $\mathbf{A}$ være en $p \times m$ matrix, så $\mathbf{A}\,\mathbf{A}' = \boldsymbol{\Sigma}$. Vi sætter V lig $\mathbf{A}$'s billedrum, i.e.

$$V = \{\mathbf{v} \in R^p | \exists \mathbf{u} \in R^m : \mathbf{v} = \mathbf{A}\,\mathbf{u}\}.$$

Da er den (affine) støtte for $\mathbf{X}$ sideunderrummet

$$\mu + V = \{\mu + \mathbf{v} | \mathbf{v} \in V\}.$$

▲

BEVIS 2.6. Forbigås ■

Envidere gælder

SÆTNING 2.10. Lad $\mathbf{X}$ være som i foregående sætning. Da er underrummet V lig den direkte sum af egenrummene svarende til de fra 0 forskellige egenværdier i Σ. ▲

BEVIS 2.7. Forbigås. ■

Endelig har vi

SÆTNING 2.11. Lad $\mathbf{X}$ være som i de foregående sætninger. Da er underrummet V lig det ortogonale komplement til nulrummet for Σ, i.e.

$$V = \{\mathbf{v} | \Sigma \mathbf{v} = \mathbf{0}\}^{\perp}$$

▲

BEVIS 2.8. Forbigås. ■

Vi illustrerer de 3 sætninger i

EKSEMPEL 2.1. Vi betragter

$$\mathbf{X} \in \mathrm{N}\left(\begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix}, \begin{bmatrix} 1 & 2 & 2 \\ 2 & 5 & 3 \\ 2 & 3 & 5 \end{bmatrix}\right) = \mathrm{N}(\mu, \Sigma).$$

Da

$$\det\left(\begin{bmatrix} 1 & 2 & 2 \\ 2 & 5 & 3 \\ 2 & 3 & 5 \end{bmatrix}\right) = 0,$$

er $\mathbf{X}$ singulært fordelt, og vi vil bestemme dens affine støtte.

Vi søger først en matrix $\mathbf{A}$, så $\mathbf{A}\,\mathbf{A}' = \mathbf{\Sigma}$. Til den ende bestemmes først $\mathbf{\Sigma}$'s egenværdier og -vektorer (normerede). De er

$$\lambda_1 = 9 \quad \wedge \quad \mathbf{p}_1 = \begin{bmatrix} \frac{1}{3} \\ \frac{2}{3} \\ \frac{2}{3} \end{bmatrix},$$

$$\lambda_2 = 2 \quad \wedge \quad \mathbf{p}_2 = \begin{bmatrix} 0 \\ \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{bmatrix},$$

$$\lambda_3 = 0 \quad \wedge \quad \mathbf{p}_3 = \begin{bmatrix} \frac{2\sqrt{2}}{3} \\ -\frac{\sqrt{2}}{6} \\ -\frac{\sqrt{2}}{6} \end{bmatrix}.$$

Følgelig er

$$\mathbf{\Sigma} = \begin{bmatrix} \frac{1}{3} & 0 & \frac{2\sqrt{2}}{3} \\ \frac{2}{3} & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{6} \\ \frac{2}{3} & -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{6} \end{bmatrix} \begin{bmatrix} 9 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{3} & \frac{2}{3} & \frac{2}{3} \\ 0 & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{2\sqrt{2}}{3} & -\frac{\sqrt{2}}{6} & -\frac{\sqrt{2}}{6} \end{bmatrix}$$

Heraf fås, at vi som $\mathbf{A}$-matrix kan vælge

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 2 & -1 & 0 \end{bmatrix}.$$

Opfattes $\mathbf{A}$ som matrix for en lineær afbildning $R^3 \to R^3$ fås, at billedrummet er

$$\begin{aligned} V &= \{\mathbf{A}\,\mathbf{u} | \mathbf{u} \in R^3\} \\ &= \{u_1\mathbf{p}_1 + u_2\mathbf{p}_2 | u_1 \in R \wedge u_2 \in R\}. \end{aligned}$$

Det ses umiddelbart, at dette også er den direkte sum af egenrummene svarende til de fra 0 forskellige egenværdier.

Nulrummet for $\mathbf{\Sigma}$ er givet ved

$$\mathbf{\Sigma}\,\mathbf{u} = \mathbf{0} \quad \Leftrightarrow \quad \mathbf{u} = t \cdot \mathbf{p}_3.$$

Heraf fås igen den samme beskrivelse af V.

Den affine støtte for $\mathbf{Y}$ er altså sideunderrummet

$$\mu + V = \left\{ \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix} + u_1 \begin{bmatrix} \frac{1}{3} \\ \frac{2}{3} \\ \frac{2}{3} \end{bmatrix} + u_2 \begin{bmatrix} 0 \\ \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{bmatrix} | u_1, u_2 \in R \right\}.$$

♦

BEMÆRKNING 2.4. Af eksemplet fremgår beviserne for sætningerne 2.9-2.11 næsten fuldstændigt. ▼

Vi formulerer nu en triviel, men nyttig sætning.

SÆTNING 2.12. Lad $\mathbf{X} \in \mathrm{N}(\mu, \Sigma)$. Da gælder

$$\mathbf{A}\,\mathbf{X} + \mathbf{b} \in \mathrm{N}(\mathbf{A}\,\mu + \mathbf{b},\ \mathbf{A}\,\Sigma\,\mathbf{A}'),$$

hvor det naturligvis forudsættes, at de anførte matrixprodukter m.v. eksisterer. ▲

BEVIS 2.9. Triviel følge af definitionen. ■

2.2.2 Uafhængighed og konturellipsoider

I dette afsnit skal vi dels give betingelser for uafhængighed af normalt fordelte stokastiske variable, og dels godtgøre, at niveaukurverne for tæthedsfunktioner er ellipsoider. Vi har først

SÆTNING 2.13. Lad

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \in \mathrm{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right).$$

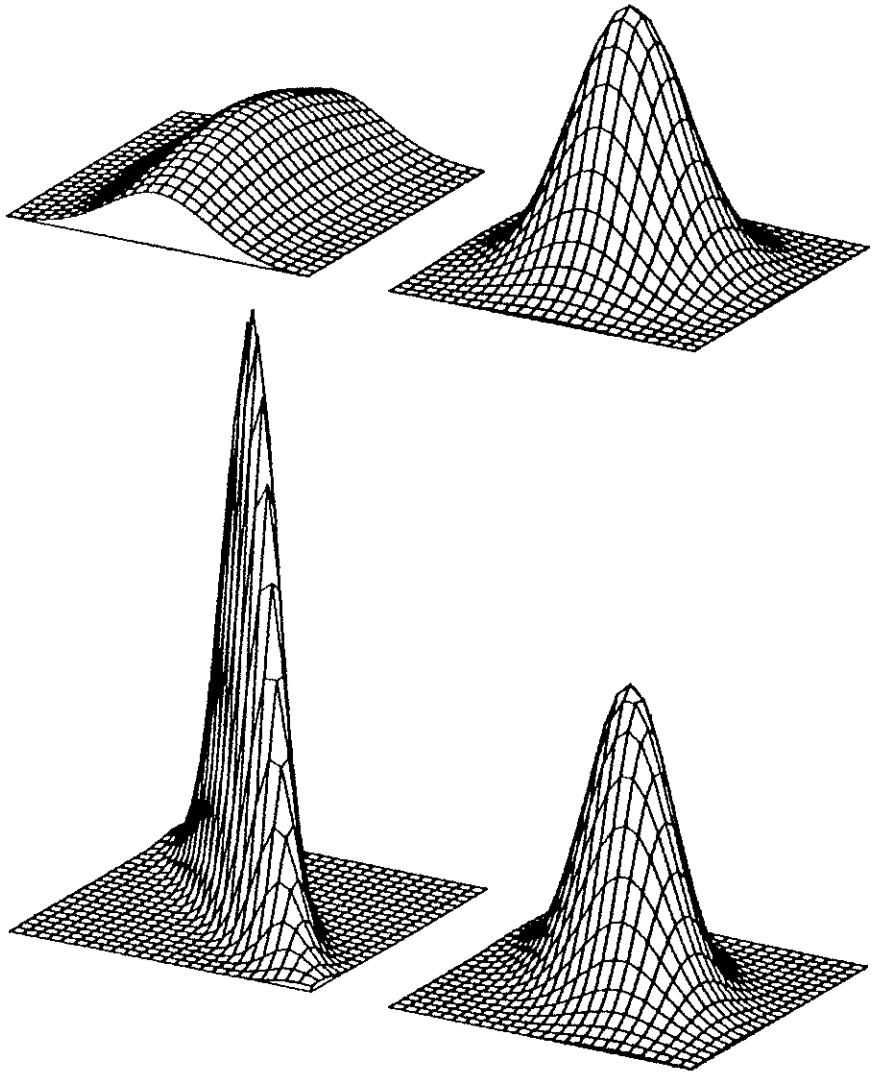
Da gælder, at

$$\mathbf{X}_i \in \mathrm{N}(\mu_i, \Sigma_{ii}),$$

og

$$\mathbf{X}_1, \mathbf{X}_2 \text{ stokastisk uafhængige} \quad \Leftrightarrow \quad \Sigma_{12} = \Sigma_{21}' = \mathbf{0},$$

hvor $\mathbf{0}$ er en matrix bestående af lutter nuller. ▲



Figur 2.1: Tæthedsfunktioner for todimensionale normalfordelinger med dispersionsmatricer $\begin{pmatrix} 1 & 0 \\ 0 & 3 \end{pmatrix}$, $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $\begin{pmatrix} 1 & -0.9 \\ -0.9 & 1 \end{pmatrix}$ og $\begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}$.

BEVIS 2.10. Første udsagn fås direkte af den foregående sætning. Det andet fås ved at godtgøre, at betingelsen $\Sigma_{12} = \mathbf{0}$ sikrer, at fordelingen bliver en produktfordeling. ■

Af sætningen følger specielt, at komponenterne i en vektor $\mathbf{X} \in \mathrm{N}(\mu, \Sigma)$ er stokastisk uafhængige, netop hvis Σ er en diagonalmatrix. Vi skal nu vise, at uafhængigheden blot er et spørgsmål om at vælge et passende koordinatsystem.

Lad $\mathbf{X} \in \mathrm{N}(\mu, \Sigma)$ og lad Σ have de ortonormerede egenvektorer $\mathbf{p}_1, \ldots, \mathbf{p}_n$. Vi betragter nu koordinatsystemet, der har begyndelsespunkt i μ og vektorerne $\mathbf{p}_1, \ldots, \mathbf{p}_n$ som basisvektorer. Koordinaterne i dette system benævnes $\mathbf{y}$.

Sættes

$$\mathbf{P} = (\mathbf{p}_1, \ldots, \mathbf{p}_n),$$

har vi følgende sammenhæng mellem de oprindelige koordinater $\mathbf{x}$ og de nye koordinater $\mathbf{y}$ for et vilkårligt punkt $\in R^n$.

$$\mathbf{y} = \mathbf{P}'(\mathbf{x} - \mu) \quad \Leftrightarrow \quad \mathbf{x} = \mathbf{P}\,\mathbf{y} + \mu,$$

jvf. p. 12.

NB. Ovenstående relation er en relation mellem **koordinater** for en **fast** vektor betragtet i 2 koordinatsystemer.

Lader vi i overensstemmelse hermed $\mathbf{Y}$ være de nye koordinater for $\mathbf{X}$ har vi

SÆTNING 2.14. Lad $\mathbf{X} \in \mathrm{N}(\mu, \Sigma)$ og lad $\mathbf{Y}$ være som ovenfor. Da gælder

$$\mathbf{Y} \in \mathrm{N}(\mathbf{0}, \Lambda),$$

hvor Λ er en diagomalmatrix med Σ's egenværdier i diagonalen. ▲

BEVIS 2.11. Følger af sætning 2.2.1 og sætning 1.10 ■

BEMÆRKNING 2.5. Ved at foretage en translation og en drejning (eller spejling) af det oprindelige koordinatsystem har vi opnået, at dispersionsmatricen er en diagonalmatrix, i.e. at komponenterne i den stokastiske vektor er ukorrelerede og dermed uafhængige. ▼

Ved at foretage en reskalering af akserne, kan vi endda opnå, at dispersionsmatricen får lutter nuller eller ettaller i diagonalen. Betragter vi nemlig basisvektorerne

$$c_1\mathbf{p}_1, \ldots, c_n\mathbf{p}_n,$$

hvor

$$c_i = \begin{cases} \frac{1}{\sqrt{\lambda_i}} & \text{hvis } \lambda_i > 0 \\ 1 & \text{hvis } \lambda_i = 0 \end{cases},$$

jvf. p. 31, og kaldes koordinaterne i dette system $\mathbf{z}$, fås sammenhængen

$$\mathbf{z} = \mathbf{C}'\mathbf{P}'(\mathbf{x} - \mu) = (\mathbf{P}\,\mathbf{C})'(\mathbf{x} - \mu),$$

hvor $\mathbf{C} = \text{diag}(c_1, \ldots, c_n)$.

Sætter vi $\mathbf{z}$-koordinaterne for $\mathbf{X}$ lig $\mathbf{Z}$ fås derfor

$$\mathbf{Z} = \text{N}(\mathbf{0}, \mathbf{E}),$$

hvor

$$\mathbf{E} = (\mathbf{P}\,\mathbf{C})'\mathbf{\Sigma}\,\mathbf{P}\,\mathbf{C} = \mathbf{C}'\mathbf{P}'\mathbf{\Sigma}\,\mathbf{P}\,\mathbf{C} = \mathbf{C}'\mathbf{\Lambda}\,\mathbf{C}$$

har nuller eller ettaller i diagonalen.

Overgangen til disse nye baser har en tæt tilknytning til niveaukurverne for tæthedsfunktion for den normale fordeling.

Tætheden for et $\mathbf{X} \in \text{N}(\mu, \mathbf{\Sigma})$ er som tidligere anført

$$\begin{aligned} \text{f}(\mathbf{x}) &= k \cdot \exp(-\frac{1}{2}(\mathbf{x} - \mu)'\mathbf{\Sigma}^{-1}(\mathbf{x} - \mu)) \\ &= k \cdot \exp(-\frac{1}{2}(||\mathbf{x} - \mu||)^2). \end{aligned}$$

Derfor er

$$\text{f}(\mathbf{x}) = k_1 \quad \Leftrightarrow \quad (\mathbf{x} - \mu)'\mathbf{\Sigma}^{-1}(\mathbf{x} - \mu) = c,$$

hvor k_1 og c er konstanter. Da $\mathbf{\Sigma}^{-1}$, er positivt definit er niveaukurverne

$$E_c = \{\mathbf{x} | \text{f}(\mathbf{x}) = k_1\}$$

derfor **ellisoider**, jvf. p. 40. Af sætning 1.19 fremgår endvidere, at hovedakserne i disse ellipsoider netop er egenvektorer til Σ^{-1}, men ifølge sætning 1.12 er disse også egenvektorer til Σ. I de nye koordinater bliver tæthederne derfor

$$g(\mathbf{y}) = k \cdot \exp(-\frac{1}{2}\Sigma\frac{1}{\lambda_i}y_i^2),$$

hvor λ_i er den i'te egenværdi til Σ, og

$$h(\mathbf{z}) = k_1 \cdot \exp(-\frac{1}{2}\Sigma z_i^2).$$

Ellipsoiderne E_i kaldes ofte **kontur-ellipsoider**.

Ved hjælp af ovenstående betragtninger får vi

SÆTNING 2.15. Lad $\mathbf{P}$ og $\mathbf{C}$ være som ovenfor. Da gælder

$$(\mathbf{X} - \mu)'(\mathbf{P}\,\mathbf{C})(\mathbf{P}\,\mathbf{C})'(\mathbf{X} - \mu) \in \chi^2(\text{rg}\,\Sigma).$$

Hvis specielt Σ har fuld rang p gælder

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

▲

BEVIS 2.12. $(\mathbf{X} - \mu)'(\mathbf{P}\,\mathbf{C})(\mathbf{P}\,\mathbf{C})'(\mathbf{X} - \mu) = \mathbf{Z}'\mathbf{Z} = \Sigma\delta_i Z_i^2,$

hvor $\delta_i = 1$ hvis $\lambda_i \neq 0$ og lig 0 ellers.

Da de ikke udartede komponenter i $\mathbf{Z}$ er stokastisk uafhængige og N(0,1)-fordelte følger resultatet umiddelbart. Den sidste bemærkning følger af, at

$$\mathbf{P}\,\mathbf{C}(\mathbf{P}\,\mathbf{C})' = \mathbf{P}\,\mathbf{C}\,\mathbf{C}'P' = \mathbf{P}\,\mathbf{\Lambda}^{-1}\mathbf{P}' = \mathbf{\Sigma}^{-1}$$

■

BEMÆRKNING 2.6. Resultatet i sætningen er, at sandsynligheden for at få et udfald inden for en kontur-ellipsoide kan beregnes ved hjælp af en χ^2-fordeling. ▼

Disse begreber vil blive eksemplificeret i eksempel 2.3, hvor vi betragter den todimensionale fordeling.

2.2.3 Betingede fordelinger

Vi betragter i dette afsnit en spaltning af en stokastisk variabel $\mathbf{X} \in \mathrm{N}_p(\mu, \Sigma)$, nemlig

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}; \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}; \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

Vi har da

SÆTNING 2.16. Hvis $\mathbf{X}_2$ er regulært fordelt, i.e. hvis Σ_{22} har fuld rang, da er den betingede fordeling af $\mathbf{X}_1$ givet $\mathbf{X}_2 = \mathbf{x}_2$ igen en normal fordeling, og der gælder

$$\begin{aligned} \mathrm{E}(\mathbf{X}_1|\mathbf{X}_2 = \mathbf{x}_2) &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \mu_2) \\ \mathrm{D}(\mathbf{X}_1|\mathbf{X}_2 = \mathbf{x}_2) &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \end{aligned}$$

Hvis Σ_{22} ikke har fuld rang, er den betingede fordeling stadig normal, og Σ_{22}^{-1} skal i ovenstående formler blot erstattes med en generaliseret invers Σ_{22}^{-}. ▲

BEVIS 2.13. Beviset er teknisk og forbigås, jvf. dog afsnit 2.2.5. ■

BEMÆRKNING 2.7. Det fremgår, at den betingede varians er uafhængig af $\mathbf{x}_2$. Dette resultat gælder selvsagt ikke for vilkårlige fordelinger, men er specielt for den normale fordeling. Endvidere fremgår, at den betingede middelværdi er en affin funktion af $\mathbf{x}_2$, jvf. betragtningerne i afsnit 2.3.3. ▼

Vi skal ikke her diskutere den nærmere betydning af sætningnen , men henvise til eksemplerne i afsnit 2.2.5.

2.2.4 Reproduktivitetssætning og central grænseværdisætning

I analogi til reproduktivitetssætningnen for den endimensionale normale fordeling har vi

SÆTNING 2.17. (**Reproduktivitetssætning**). Lad $\mathbf{X}_1, \ldots, \mathbf{X}_k$ være uafhængige, og lad $\mathbf{X}_i \in \mathrm{N}(\mu_i, \Sigma_i)$.

Da er

$$\sum_{i=1}^{k} \mathbf{X}_i \in \mathrm{N}\left(\sum_{i=1}^{k} \mu_i, \sum_{i=1}^{k} \Sigma_i\right).$$

▲

BEVIS 2.14. Forbigås. ■

Som i det endimensionale tilfælde gælder der også centrale grænseværdisætninger, i.e. at summer af uafhængige flerdimensionale stokastiske variable under generelle forudsætninger er asymptotisk normalt fordelte. Vi anfører en analog til Lindeberq-Levy's sætning.

SÆTNING 2.18. **(Central grænseværdisætning).** Lad de uafhængige og identisk fordelte variable $X_l, \ldots, X_n, \ldots$ have endeligt første og andet moment

$$\mu = \mathrm{E}(\mathbf{X}_i), \Sigma = \mathrm{D}(\mathbf{X}_i).$$

Da vil - med $\bar{\mathbf{X}}_n = \frac{1}{n}(\mathbf{X}_1 + \cdots + \mathbf{X}_n)$-

$$\sqrt{n}(\bar{\mathbf{X}}_n - \mu)$$

have en $\mathrm{N}(\mathbf{0}, \Sigma)$-fordeling som grænsefordeling, og vi siger, at $\bar{\mathbf{X}}_n$ er asymptotisk $\mathrm{N}(\mu, \frac{1}{n}\Sigma)$-fordelt. ▲

BEVIS 2.15. Såvel denne som den foregående sætning kan vises ud fra de tilsvarende endimensionale sætninger ved dels at bruge en sætning, som karakteriserer den flerdimensionale fordeling (en flerdimensional variabel er normalt fordelt, netop hvis enhver linearkombination af dens komponenter er (endimensionalt) normalt fordelt), og dels en sætning, som karakteriserer en flerdimensional grænsefordeling ved grænsefordelinger af linearkombinationer af komponenterne (koordinaterne). Dette ligger dog ud over denne fremstillings rammer, og for nærmere detaljer må den interesserede læser henvises til litteraturen, e.g. [28], afsnit 2c.5. ■

2.2.5 Estimation af parametre i en flerdimensional normal fordeling

Vi betragter en række observationer $\mathbf{X}_1, \ldots, \mathbf{X}_n$, der antages uafhængige og identisk $\mathrm{N}_p(\mu, \Sigma)$-fordelte. Vi antager, at der er flere observationer end dimensionen angiver, d.v.s. at $n > p$. Vi skal i dette afsnit anføre skøn over parametrene μ og Σ.

Vi indfører betegnelserne

$$\mathbf{X}_i = \begin{bmatrix} X_{1i} \\ \vdots \\ X_{pi} \end{bmatrix}$$

$$\bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i = \begin{bmatrix} \bar{X}_1 \\ \vdots \\ \bar{X}_p \end{bmatrix}$$

$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})' = \frac{1}{n-1}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i' - \frac{n}{n-1}\bar{\mathbf{X}}\,\bar{\mathbf{X}}'.$$

Betragter vi datamatricen

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1' \\ \vdots \\ \mathbf{X}_n' \end{bmatrix} = \begin{bmatrix} X_{11} & \cdots & X_{p1} \\ \vdots & & \vdots \\ X_{1n} & \cdots & X_{pn} \end{bmatrix},$$

hvor den i'te række netop er den i'te observation, kan vi også skrive

$$(n-1)\mathbf{S} = \sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})' = \mathbf{X}'\mathbf{X} - n\bar{\mathbf{X}}\bar{\mathbf{X}}'.$$

Med disse betegnelser er vi i stand til at formulere

SÆTNING 2.19. Lad situationen være som ovenfor beskrevet. Da er maximum likelihood estimatorerne for μ og Σ lig

$$\hat{\mu} = \bar{\mathbf{X}}$$

$$\hat{\Sigma} = \frac{n-1}{n}\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})'.$$

$\hat{\mu}$ er et centralt skøn over μ, og $\mathbf{S}$ er et centralt skøn over Σ. ▲

BEVIS 2.16. Forbigås, se f.eks. [3], kapitel 3. ■

BEMÆRKNING 2.8. Da den **empiriske dispersionsmatrix** $\mathbf{S}$ er et centralt skøn over Σ, og da den kun afviger fra maximun likelihood estimatoren med faktoren $\frac{n}{n-1}$, foretrækkes $\mathbf{S}$ ofte som skøn. Man vil da også hyppigt anvende betegnelsen $\hat{\Sigma}$ om $\mathbf{S}$. Man

må derfor i konkrete fremstillinger være opmærksom på, hvad et udtryk som $\hat{\Sigma}$ præcist dækker over.

Fordelingen af $\hat{\mu}$ fås trivielt ved hjælp af sætning 2.2.4. Der gælder

$$\hat{\mu} = \bar{X} \in \mathrm{N}_p(\mu, \frac{1}{n}\Sigma).$$

Fordelingen af S er mere kompliceret, Den er anført i afsnit 2.5.

Vi anfører et eksempel på estimation af parametrene i næste afsnit. ▼

2.2.6 Den todimensionale normale fordeling

Vi specialiserer nu de tidligere resultater til to dimensioner.

Lad $\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ være normalt fordelt (μ, Σ), hvor

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

og

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

Da

$$\det(\Sigma) = \sigma_1^2\sigma_2^2 - \sigma_{12}^2$$

er, såfremt $\det(\Sigma) \neq 0$,

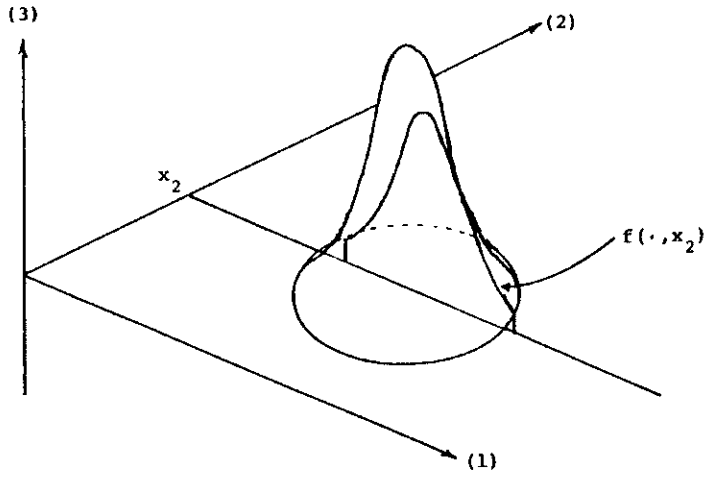
$$\Sigma^{-1} = \frac{1}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2} \begin{bmatrix} \sigma_2^2 & -\sigma_{12} \\ -\sigma_{12} & \sigma_1^2 \end{bmatrix}.$$

Indføres korrelationskoefficienten ρ

$$\rho = \frac{\sigma_{12}}{\sigma_1\,\sigma_2},$$

fås

$$\Sigma^{-1} = \frac{1}{1-\rho^2} \begin{bmatrix} \frac{1}{\sigma_1^2} & \frac{-\rho}{\sigma_1\sigma_2} \\ \frac{-\rho}{\sigma_1\sigma_2} & \frac{1}{\sigma_2^2} \end{bmatrix},$$



Figur 2.2: Tæthed for to-dimensional normal fordeling

og tætheden bliver

$$\mathrm{f}(x_1, x_2) = \frac{1}{2\pi} \frac{1}{\sigma_1 \sigma_2 \sqrt{1-\rho^2}} \exp\left[-\frac{1}{2} \frac{1}{1-\rho^2} \left\{ \left[\frac{x_1 - \mu_1}{\sigma_1} \right]^2 - 2\rho \frac{x_1 - \mu_1}{\sigma_1} \frac{x_2 - \mu_2}{\sigma_2} + \left[\frac{x_2 - \mu_2}{\sigma_2} \right]^2 \right\} \right].$$

Grafen er skitseret i fig. 2.2 Det ses umiddelbart, at der er tale om en produktfordeling, i.e. at X_1 og X_2 er stokastisk uafhængige, netop hvis $\rho = 0$, d.v.s. netop hvis Σ er en diagonalmatrix.

Den betingede tæthed for X_1 givet $X_2 = x_2$ er proportional med skæringskurven mellem planen gennem $(0, x_2, 0)$ parallel med $(1) - (3)$ planen. Kaldes tætheden g har vi altså

$$\mathrm{g}(\cdot) = c\mathrm{f}(\cdot, x_2),$$

hvor c er en normeringskonstant. Vi har

$$\begin{aligned} \mathrm{g}(x_1) &= k_1 \cdot \exp\left[-\frac{1}{2}\frac{1}{1-\rho^2}\left\{\left[\frac{x_1-\mu_1}{\sigma_1}\right]^2 - 2\rho\frac{x_1-\mu_1}{\sigma_1}\frac{x_2-\mu_2}{\sigma_2}\right\}\right] \\ &= k_2 \cdot \exp\left[-\frac{1}{2}\frac{1}{1-\rho^2}\left[\frac{x_1-\mu_1}{\sigma_1} - \rho\frac{x_2-\mu_2}{\sigma_2}\right]^2\right] \\ &= k_3 \cdot \exp\left[-\frac{1}{2}\frac{1}{\sigma_1^2(1-\rho^2)}\left(x_1-\mu_1-\rho\frac{\sigma_1}{\sigma_2}(x_2-\mu_2)\right)^2\right] \\ &= k_3 \cdot \exp\left[-\frac{1}{2\gamma^2}(x_1-\xi_1)^2\right]. \end{aligned}$$

Vi har her ikke holdt regnskab med led der kun indeholder x_2. De er gået ud som diverse konstanter. Af slutresultatet aflæses, at den betingede fordeling er normal, og at

$$k_3 = \frac{1}{\sqrt{2\pi}\sigma_1\sqrt{1-\rho^2}},$$

samt endelig, at

$$\mathrm{E}(X_1|X_2=x_2) = \xi_1 = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2-\mu_2)$$

og

$$\mathrm{V}(X_1|X_2=x_2) = \gamma^2 = \sigma_1^2(1-\rho^2).$$

Vi har bevist resultatet i sætning 2.16 i tilfældet $n = 2$. Det bemærkes, at den betingede middelværdi afhænger lineært (rettere affint) af x_2, og at den betingede varians er uafhængig af x_2. Endvidere er

$$\mathrm{V}(X_1|X_2=x_2) \leq \mathrm{V}(X_1),$$

og den **kvadrerede korrelationskoefficient angiver variansreduktionen, d.v.s. den brøkdel af X_1's varians, der forklares af X_2**, idet

$$\rho^2 = \frac{\mathrm{V}(X_1) - \mathrm{V}(X_1|X_2=x_2)}{\mathrm{V}(X_1)}.$$

I det følgende eksempel betragter vi et numerisk eksempel, der også involverer et estimationproblem.

EKSEMPEL 2.2. I nedenstående tabel ar anført sammenhørende målinger af luftens indhold af svævestøv målt i $\mu\frac{g}{m^3}$. Ved målingerne er anvendt to forskellige måleprincipper, nemlig et sværtningsprincip (med et såkaldt OECD-apparat) og et vejningsprincip (med en såkaldt High Volume Sampler). Grunden til, at der kan optræde store afvigelser er bl.a. , at målinger ved sværtningsprincippet er meget følsomme overfor svævestøvs eventuelle afvigelse fra "normalstøv". Således vil et stort indhold af kalkstøv i luften kunne bevirke, at målingerne bliver systematisk for små.

Metode	I	2	5	15	16	16	19	26	24	16	36
	II	2	12	4	21	41	14	31	29	31	8
	I	39	42	44	40	42	42	50	51	58	64
	II	30	44	26	60	34	34	14	41	58	47

Vi opfatter ovenstående data som realiserede udfald af uafhængige, identisk fordelte stokastiske variable

$$\begin{bmatrix} X_1 \\ Y_1 \end{bmatrix}, \ldots, \begin{bmatrix} X_{20} \\ Y_{20} \end{bmatrix}.$$

Vi vil undersøge om fordelingen kan antages at være normal med parametre (μ, Σ). Såfremt fordelingen er normal, finder vi skønnene

$$\hat{\mu} = \begin{bmatrix} \hat{\mu_1} \\ \hat{\mu_2} \end{bmatrix} = \begin{bmatrix} \bar{X} \\ \bar{Y} \end{bmatrix} = \begin{bmatrix} 32.35 \\ 29.05 \end{bmatrix},$$

og

$$\hat{\Sigma} = \begin{bmatrix} \hat{\sigma_1^2} & \hat{\sigma_{12}} \\ \hat{\sigma_{12}} & \hat{\sigma_2^2} \end{bmatrix} = \begin{bmatrix} S_1^2 & S_{12} \\ S_{12} & S_2^2 \end{bmatrix} = \begin{bmatrix} 311 & 182 \\ 182 & 279 \end{bmatrix},$$

hvor $\hat{\Sigma}$ er det centrale skøn over Σ. Specielt er

$$S_{12} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}).$$

Vi vil altså undersøge, om observationerne kan tænkes at komme fra en normal fordeling med parametre $(\hat{\mu}, \hat{\Sigma})$. Til den ende bestemmer vi først konturellipserne. Egenværdierne og egenvektorerne for $\hat{\Sigma}$ er

$$\hat{\lambda_1} = 477.613 \quad \text{og} \quad \hat{\mathbf{p}_1} = \begin{bmatrix} 0.736 \\ 0.678 \end{bmatrix}$$

og

$$\hat{\lambda}_2 = 112.676 \quad \text{og} \quad \hat{\mathbf{p}}_2 = \begin{bmatrix} -0.678 \\ 0.736 \end{bmatrix}.$$

Vælger vi koordinatsystemet med centrum i $\hat{\mu}$ og med $\mathbf{p}_1$ og $\mathbf{p}_2$ som basisvektorer, får konturellipserne ligninger af formen

$$\frac{z_1^2}{\hat{\lambda}_1} + \frac{z_2^2}{\hat{\lambda}_2} = c,$$

eller

$$\frac{z_1^2}{477.613} + \frac{z_2^2}{112.676} = c,$$

hvor de nye koordinater er givet ved

$$\mathbf{P}\,\mathbf{z} = (\mathbf{p}_1\mathbf{p}_2)\mathbf{z} = \mathbf{x} - \hat{\mu}.$$

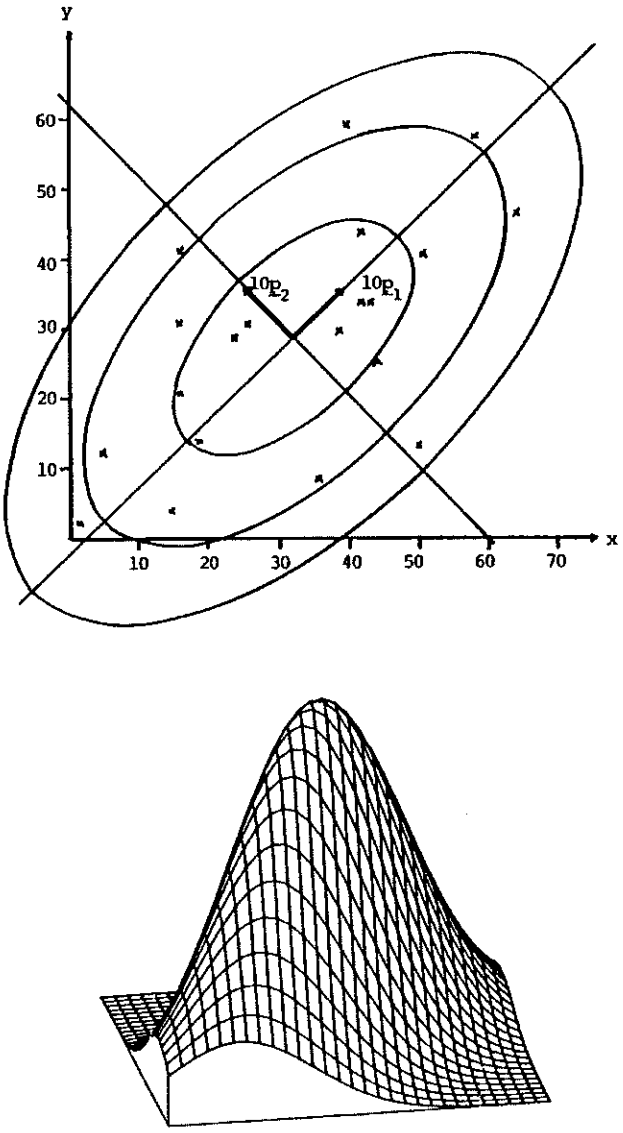
I figur 2.2 er der indtegnet observationerne og 3 konturellipser svarende til c-værdierne $c_1 = \chi^2(2)_{0.40}$, $c_2 = \chi^2(2)_{0.80}$ og $c_3 = \chi^2(2)_{0.95} = 5.99$.
Dette bevirker, ifølge sætning 2.15, at der i normalfordelingen med parametre $(\hat{\mu}, \hat{\Sigma})$ er sandsynlighederne 40%, 80% og 95% for at få observationer indenfor den inderste, den mellemste og den yderste ellipse. For områderne mellem ellipserne, resp. uden for disse, fås da sandsynlighederne 40%, 40%, 15% og 5%. Disse tal kan sammenlignes med de tilsvarende observerede relative hyppigheder, som er 40%, 30%, 30% og 0%. Overensstemmelsen er, om ikke overvældende, så dog acceptabel.

Hvis man ønsker et mere præcist resultat kan man udføre et χ^2-test. Det vil da være rimeligt at foretage en yderligere inddeling af planen efter egenvektorernes retninger. I det aktuelle tilfælde får vi da 4×4 områder med estimerede sandsynligheder på 10%, 10%, 3.75% og 1.25%. Man kan da beregne den sædvanlige χ^2-teststørrelse

$$\sum \frac{(\text{obs.} - \text{forv.})^2}{\text{forv.}}$$

og sammenligne med en $\chi^2(n-6)$-fordeling (vi har estimeret 5 parametre). I det konkrete tilfælde er der måske lige lovligt få observationer til at kunne foretage denne analyse.
Korrelationskoefficienten estimeres til

$$\hat{\rho} = \frac{182}{\sqrt{311 \cdot 279}} = 0.62,$$



Figur 2.3: Estimerede konturellipser og estimeret tæthedsfunktion svarende til data i eksempel 2.2

hvorfor de betingede varianser skønnes til

$$\begin{aligned}\hat{\mathrm{V}}(X|Y=y) &= 311(1-\hat{\rho}^2) = 192\\ \hat{\mathrm{V}}(Y|X=x) &= 279(1-\hat{\rho}^2) = 172.\end{aligned}$$

Vi ser, at de betingede varianser er blevet reduceret med 38% svarende til, at $\rho^2 = 0.38$. Dette at den betingede varians af f.eks. en OECD-måling for fastholdt High Volume Sampler måling er væsentlig mindre end den ubetingede varians virker vel også ganske rimeligt. Ved vi, at svævestøvsindholdet ved hjælp af en High Volume Sampler er bestemt til e.g. $2\mu\frac{g}{m^3}$, vil vi vel ikke forvente at få resultater med OECD-apparatet, der afviger væsentligt fra dette, svarende til, at vi får en beskeden betinget varians. Kendes High Volume Sampler resultatet ikke, ja da må vi forvente, at OECD-målingen vil kunne ligge i hele det naturlige variationsområde, svarende til en større, ubetinget varians. ♦

2.3 Korrelation og regression

I dette afsnit skal vi nøjere diskutere parametrenes betydning i en flerdimensional normal fordeling. Vi vil først forsøge at generalisere de egenskaber ved korrelations koefficienten, som fremgik af foregående afsnit.

2.3.1 Den partielle korrelationskoefficient

Udgangspunktet er formlen for de betingede fordelinger i en flerdimensional normal fordeling. Lad $\mathbf{X} \in \mathrm{N}_p(\mu, \Sigma)$, og lad de variable være spaltet som følger

$$\mathbf{X} = \begin{bmatrix}\mathbf{X}_1\\ \mathbf{X}_2\end{bmatrix};\quad \mu = \begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix};\quad \Sigma = \begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{bmatrix},$$

hvor $\mathbf{X}_1$ består af de m første pladser i $\mathbf{X}$ og analogt med de øvrige. Da er den betingede dispersion af $\mathbf{X}_1$ for givet $\mathbf{X}_2 = x_2$ som anført i sætning 2.16 lig

$$\mathrm{D}(\mathbf{X}_1|\mathbf{X}_2 = x_2) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

Ved den **partielle korrelationskoefficient** mellem X_i og X_j, $i,j \leq m$, for givet $\mathbf{X}_2 = \mathbf{x}_2$ forstås da korrelationen i den betingede fordeling af $\mathbf{X}_1$ givet $\mathbf{X}_2 = \mathbf{x}_2$. Den betegnes $\rho_{ij|m+1,\ldots,p}$.

Sætter vi

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \cdots & \sigma_{p1} \\ \vdots & & \vdots \\ \sigma_{1p} & \cdots & \sigma_p^2 \end{bmatrix}$$

og

$$\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{21} = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{1m} & \cdots & a_{mm} \end{bmatrix},$$

har vi altså

$$\rho_{ij|m+1,\ldots,n} = \frac{a_{ij}}{\sqrt{a_{ii}}\sqrt{a_{jj}}}.$$

Er specielt $\mathbf{X}$ tredimensionel fås med

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & \rho_{13}\sigma_1\sigma_3 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & \rho_{23}\sigma_2\sigma_3 \\ \rho_{13}\sigma_1\sigma_3 & \rho_{23}\sigma_2\sigma_3 & \sigma_3^2 \end{bmatrix},$$

at

$$\begin{aligned} &\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} \\ &= \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} - \frac{1}{\sigma_3^2}\begin{bmatrix} \rho_{13}^2\sigma_1^2\sigma_3^2 & \rho_{13}\rho_{23}\sigma_1\sigma_2\sigma_3^2 \\ \rho_{13}\rho_{23}\sigma_1\sigma_2\sigma_3^2 & \rho_{23}^2\sigma_2^2\sigma_3^2 \end{bmatrix} \\ &= \begin{bmatrix} \sigma_1^2(1-\rho_{13}^2) & \sigma_1\sigma_2(\rho_{12}-\rho_{13}\rho_{23}) \\ \sigma_1\sigma_2(\rho_{12}-\rho_{13}\rho_{23}) & \sigma_2^2(1-\rho_{23}^2) \end{bmatrix}. \end{aligned}$$

Heraf følger, at den partielle korrelationskoefficient mellem X_1 og X_2 givet X_3 er

$$\rho_{12|3} = \frac{\rho_{12} - \rho_{13}\rho_{23}}{\sqrt{(1-\rho_{13}^2)(1-\rho_{23}^2)}}.$$

For en p-dimensional vektor $\mathbf{X}$ finder vi derfor

$$\rho_{ij|k} = \frac{\rho_{ij} - \rho_{ik}\rho_{jk}}{\sqrt{(1-\rho_{ik}^2)(1-\rho_{jk}^2)}}. \qquad (**)$$

	C_3S	C_3A	BLAINE	Styrke 3	Styrke 28
C_3S	1	-0.309	0.901	0.158	0.344
C_3A	-0.309	1	0.192	0.120	-0.166
BLAINE	0.091	0.192	1	0.745	0.320
Styrke 3	0.158	0.120	0.745	1	0.464
Styrke 28	0.344	-0.168	0.320	0.464	1

Tabel 2.1: Korrelationsmatrix for 5 cementvariable.

Da man kan finde betingede fordelinger for givet $X_{m+1}, \ldots, X_p$ ved successive betingninger, kan man derfor få partielle korrelationskoefficienter af højere orden ved successiv anvendelse af (**). Eksempelvis findes

$$\rho_{ij|kl} = \frac{\rho_{ij|k} - \rho_{il|k} \cdot \rho_{jl|k}}{\sqrt{(1-\rho^2_{il|k}) \cdot (1-\rho^2_{jl|k})}},$$

hvor vi først har betinget med X_k og dernæst med X_l.

I afsnit 2.2.6 så vi, at korrelationskoefficienten er et mål for den variansreduktion vi opnår ved at betinge med den ene variabel. Da de partielle korrelationskoefficienter blot er korrelationer i betingede fordelinger, kan vi her anvende den samme tolkning. F.eks. har vi, at $\rho^2_{ij|kl}$ angiver den brøkdel af X_i's varians for fastholdt $X_k = x_k$ og $X_l = x_l$ der forklares af X_j. Det må her indskærpes, at disse tolkninger er **snævert knyttede til forudsætningen om normalitet.** I det generelle tilfælde, vil de betingede varianser afhænge af de værdier, hvormed der betinges (i.e. afhænge af x_k og x_l).

Ved **estimation af partielle korrelationer** estimerer man blot dispersionsmatricen og beregner derefter de partielle korrelationer som anført. **Hvis estimatet af dispersionsmatricen er en maksimum likelihood estimator, da vil de skøn over de partielle korrelationer, man får frem på denne måde også være maksimum likelihood estimatorer** (jvf. sætning 10 p. 2.28 i bind 1).

Vi vil nu give en illustration af begreberne i

EKSEMPEL 2.3. (Data stammer fra [27]).

I nedenstående tabel er anført korrelationskoefficienter mellem 3- og 28-døgnsstyrken for Portland Cement og cementens indhold af mineralerne C_3S (Alit, tricalciumsilikat, Ca_3SiO_5) og C_3A (Aluminat, tricalciumaluminat, $Ca_3Al_2O_6$), samt finheden (BLAINE. Korrelationerne er estimeret på basis af 51 sammenhørende observationer.

Det skal yderligere oplyses, at C_3S udgør ca. $35-60\%$ af almindelige portlandklinker og C_3A ca. $5-18\%$ af klinkerne. BLAINE-tallet er et mål for den specifikke overflade, således at et stort BLAINE tal svarer til en meget finmalet cement.

Vi vil især interessere os for sammenhængen mellem C_3A indholdet i klinkerne og styrkerne. Det er almindeligt anerkendt jvf. nedenstående figur, at et stort C_3A-indhold

	C_3S	C_3A	Styrke 3	Styrke 28
C_3S	1	-0.333	0.137	0.333
C_3A	-0.333	1	-0.035	-0.246
Styrke 3	0.137	-0.035	1	0.358
Styrke 28	0.333	-0.246	0.358	1

Tabel 2.2: Korrelationsmatrix for 4 cementvariable for fastholdt BLAINE-tal.

giver en større 3-døgnsstyrke, hvilket også er i overensstemmelse med $\hat{\rho}_{C_3A,Strk3} = 0.120$. Problemet er, om denne større 3-døgnsstyrke for cementer med højt C_3A-indhold alene skyldes C_3A's større hydratiseringsgrad (jo hurtigere vandet går i forbindelse med cementen, jo hurtigere vil den selvsagt reagere og få øget styrke). C_3A's langt større hydratisering efter 3 døgn fremgår af figur 2.4(c), og hydratiseringsgradens indflydelse på styrken er **skitseret** i figur 2.4(d).

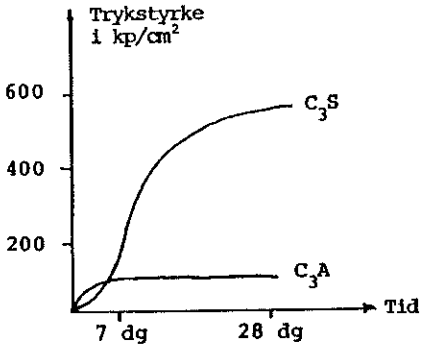
Ser vi på korrelationsmatricen fremgår imidlertid også, at C_3A-indholdet er positivt korreleret med BLAINE-tallet, d.v.s. at cementer med meget C_3A som regel vil være ganske fintmalede, og som det fremgår af figur 2.4(b) skulle dette også medvirke til en øgning af styrkerne.

Endelig bemærker vi, at 28-døgnsstyrken er svagt negativt korreleret med C_3A-indholdet. Dette virker ikke urimeligt, når vi betragter C_3S's og C_3A's styrkeudvikling med tiden, som f.eks. anført i figur 2.4(a), selv om den øgede finhed (for en C_3A-rig cement) også skulle afspejles i 28 døgnsstyrkerne, jvf. figur 2.4(b).

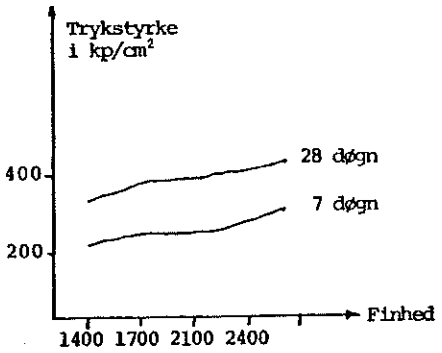
For at adskille disse forskellige eqenskaber ved C_3A fra de effekter, der opstår ved at C_3A-rig cement åbenbart er lettere at male og derfor som regel optræder i en lidt mere fintmalet form, finder vi de betingede korrelationer for fastholdt BLAINE-tal. De bliver, se tabel 2.3. Vi ser, at den partielle korrelationskoefficient mellem styrke 3 og C_3A for givet finhed er negativ (hvorimod den ubetingede korrelationskoefficient var positiv). Dette indebærer, at vi for fastholdt finhed må forvente, at cementer med højt C_3A indhold vil tendere til at have lavere styrker. Dette kunne indicere, at de store 3-døgnsstyrker for cementer med højt C_3A-indhold måske snarere skyldes, at disse cementer har højt BLAINE-tal (d.v.s. at de åbenbart males noget lettere) end, at C_3A hydratiserer hurtigt!

Vi ser en tilsvarende effekt på korrelationen mellem C_3A og Styrke 28. Her er den ubetingede korrelation -0.168 og den partielle for fastholdt BLAINE-tal formindsket til -0.246. ♦

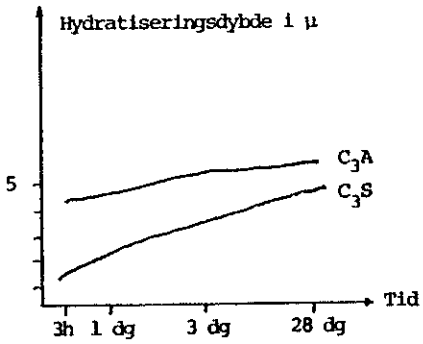
BEMÆRKNING 2.9. Ovenstående eksempel viser, at man må være meget varsom med fortolkninger af korrelationskoefficienter. Det vil f.eks. være direkte misvisende at udtale, at et højt C_3A indhold sikrer en stor 3-døgnstyrke. For det første kan man ikke slutte noget om eventuelle årsagssammenhænge ved at vurdere korrelationer. Det man kan udtale i første omgang er, at der er en tendens til, at højt C_3A-indhold og høj 3-



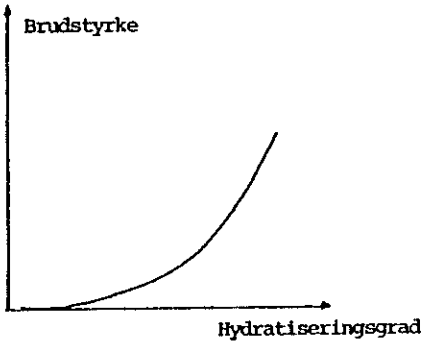
(a) Trykbrudsstyrker ved stuetemperatur af pastaer af C_3S og C_3A lagret til forskellig tid. (Efter [21]).



(b) Trykstyrker for forskellige finheder af cementen. (efter [21]).



(c) Hydratiseringsgrad for cementmineraler i forhold til tid (efter [21]).



(d) Sammenhæng mellem hydratiseringsgrad målt ved såkaldt korrigeret glødetab og brudstyrke (efter [21]).

Figur 2.4:

døgnsstyrke optræder samtidigt. Dette kan meget vel skyldes at de begge afhænger af en ukendt 3.die faktor uden, at der er tale om nogen direkte sammenhæng mellem de to variable. For det andet ser vi også, at en overgang til partielle korrelationer kan give et fortegnsskifte svarende til en effekt,der er modsat den,man får ved den direkte analyse. Dette skyldes en samvariation med den tredie faktor (her BLAINE), som forstyrrer billedet. ▼

I mange situationer vil det være ønskeligt at kunne **teste**, hvorvidt en korrelationskoefficient kan antages at være 0. Man kan da benytte

SÆTNING 2.20. Lad $R = R_{ij|m+1\ldots p}$ være den empiriske partielle korrelationskoefficient mellem X_i og X_j for givet $X_{m+1,\ldots,X_p}$. Den forudsættes beregnet ud fra det centrale skøn over dispersionsmatricen, og på grundlag af n observationer. Da er

$$\frac{R}{\sqrt{1-R^2}}\sqrt{n-2-(p-m)} \in \mathrm{t}(n-2-(p-m)),$$

hvis $\rho_{ij|m+1,\ldots,p} = 0$. ▲

BEVIS 2.17. Forbigås. ■

BEMÆRKNING 2.10. Tallet $(p-m)$ angiver det antal variable, der fastholdes. Antallet af frihedsgrader er derfor blot lig antallet af observationer minus 2 minus antal fastholdte variable. **Sætningen er også gyldig, hvis $p - m = 0$, d.v.s. hvis der er tale om en ubetinget korrelationskoefficient.** ▼

Vi fortsætter eksempel 2.3 i

EKSEMPEL 2.4. Lad os undersøge, om den fundne værdi af $r_{24|3}$ er signifikant forskellig fra 0. Vi finder med $r_{24|3} = R$:

$$\begin{aligned}\frac{R}{\sqrt{1-R^2}}\sqrt{n-2-(p-m)} &= \frac{-0.035}{\sqrt{1-0.035^2}}\cdot\sqrt{51-2-(5-4)} \\ &= -0.243 = \mathrm{t}(48)_{40\%}.\end{aligned}$$

En hypotese om, at $\rho_{24|3}$ er 0 vil derfor blive accepteret ved et test på niveau α for $\alpha < 80\%$. ♦

Hvis vi ønsker at **teste andre værdier for** ρ eller at bestemme **konfidensintervaller**, kan vi benytte

SÆTNING 2.21. Lad situationen være som i foregående sætning. Vi betragter hypotesen

$$H_0 : \rho_{ij|m+1,\ldots,p} = \rho_0$$

mod

$$H_1 : \rho_{ij|m+1,\ldots,p} \neq \rho_0.$$

Vi sætter

$$Z = \frac{1}{2} \log \frac{1 + R_{ij|m+1,\ldots,p}}{1 - R_{ij|m+1,\ldots,p}}$$

og

$$z_0 = \frac{1}{2} \log \frac{1 + \rho_0}{1 - \rho_0}.$$

Under H_0 vil da

$$(Z - z_0) \cdot \sqrt{n - (p - m) - 3} \quad \text{app.} \in \mathrm{N}(0, 1).$$

▲

BEVIS 2.18. Forbigås. ■

EKSEMPEL 2.5. Lad os eksempelvis bestemme et 95% konfidensinterval for $\rho_{24|3}$ i eksempel 2.4. Nu er

$$\begin{aligned} & \mathrm{P} \quad \{-1.96 < (Z - z) \cdot \sqrt{51 - (5 - 4) - 3} < 1.96\} \simeq 95\% \\ \Leftrightarrow \quad & \mathrm{P}\{-1.96 - 6.86Z < -6.86z < 1.96 - 6.86Z\} \simeq 95\% \\ \Leftrightarrow \quad & \mathrm{P}\{Z - 0.29 < z < Z + 0.29\} \simeq 95\%. \end{aligned}$$

Sammenhængen mellem z og $\rho_{24|3} = \rho$ er

$$z = \frac{1}{2} \log \frac{1 + \rho}{1 - \rho} \quad \Leftrightarrow \quad \rho = \frac{e^{2z} - 1}{e^{2z} + 1}$$

Den observerede værdi af Z er

$$Z = \frac{1}{2} \log \frac{1 - 0.035}{1 + 0.035} = -0.03501.$$

Grænserne for z bliver

$$[-0.3250, 0.2549].$$

De tilsvarende grænser for $\rho_{25|4}$ er

$$\left[\frac{e^{-0.6500}-1}{e^{-0.6500}+1}, \frac{e^{0.5098}-1}{e^{0.5098}+1}\right] = [-0.31, 0.25].$$

♦

2.3.2 Den multiple korrelationskoefficient

De partielle korrelationskoefficienter er en mulig udvidelse af korrelationen mellem 2 variable. Disse partielle størrelser sigter mere på at beskrive samvariationen mellem to variable. Vi vil nu forlade dette sigte og i stedet betragte formlen (p. 79)

$$\rho^2 = \frac{\mathrm{V}(X_1) - \mathrm{V}(X_1|X_2 = x_2)}{\mathrm{V}(X_1)},$$

Som giver korrelationskoefficientens variationsreducerende betydning. Dette ønsker vi nu at generalisere. Vi betragter igen opspaltningen af den p-dimensionale normalt fordelte vektor $\mathbf{X}$ i en m-dimensional vektor $\mathbf{X}_1$ og en $(p-m)$-dimensional vektor $\mathbf{X}_2$, og den deraf affødte opspaltning af parametrene, d.v.s.

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}; \quad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}; \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

Vi definerer da den multiple **korrelationskoefficient** mellem X_i, $i = 1, \ldots, m$ og $\mathbf{X}_2$ som den **maksimale korrelation** mellem X_i og en linearkombination af $\mathbf{X}_2$'s elementer. Den betegnes $\rho_{i|m+1,\ldots,p}$.

Det kan vises, at den "optimale" linearkombination af $\mathbf{X}_2$'s elementer er

$$\beta_i'\mathbf{X}_2 = (\Sigma_{12}\Sigma_{22}^{-1})_i\mathbf{X}_2,$$

hvor β_i' er den i'te række i matricen $\Sigma_{12}\Sigma_{22}^{-1}$. Det er denne matrix, der forekommer i udtrykket for den betingede middelværdi af $\mathbf{X}_1$ for givet $\mathbf{X}_2$. Denne er jo

$$\mathrm{E}(\mathbf{X}_1|\mathbf{X}_2 = \mathbf{x}_2) = \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \mu_2) = \begin{bmatrix} \beta_1' \\ \vdots \\ \beta_m' \end{bmatrix}(\mathbf{x}_2 - \mu_2).$$

Det kan også vises, at

$$\inf_{\alpha} \mathrm{V}(X_i - \alpha' \mathbf{X}_2) = \mathrm{V}(X_i - \beta_i' \mathbf{X}_2),$$

d.v.s. den betragtede linearkombination minimaliserer variansen af $X_i - \alpha' \mathbf{X}_2$.

Vi har nu følgende vigtige

SÆTNING 2.22. Vi betragter den ovenstående situation. Lad σ_i være den i'te søjle i $\mathbf{\Sigma}_{21}$, d.v.s. σ_i' den i'te række i $\mathbf{\Sigma}_{12}$.

Da gælder

$$\rho_{i|m+1,\ldots,p} = \frac{\sqrt{\sigma_i' \mathbf{\Sigma}_{22}^{-1} \sigma_i}}{\sqrt{\sigma_{ii}}}.$$

Sættes

$$\mathbf{\Sigma}_i = \begin{bmatrix} \sigma_{ii} & \sigma_i' \\ \sigma_i & \mathbf{\Sigma}_{22} \end{bmatrix},$$

er

$$1 - \rho_{i|m+1,\ldots,p}^2 = \frac{det \mathbf{\Sigma}_i}{\sigma_{ii} det \mathbf{\Sigma}_{22}} = \frac{\mathrm{V}(X_i|\mathbf{X}_2)}{\mathrm{V}(X_i)},$$

▲

BEVIS 2.19. Beviserne for påstande inden sætningen er ret simple. Der skal blot indføres en Lagrange multiplikator og benyttes, at dispersionsmatricen er positivt semidefinit.

Postulaterne i sætningen følger ved anvendelse af formlen for den betingede dispersionsstruktur (p. 74) anvendt på $\mathbf{\Sigma}_i$ ved hjælp af matrixformlerne i afsnit 1.2.7. ■

BEMÆRKNING 2.11. Med sætningen har vi fået en række karakteristika for den multiple korrelationskoefficient, og da

$$\rho_{i|m+1,\ldots,p}^2 = \frac{\mathrm{V}(X_i) - \mathrm{V}(X_i|\mathbf{X}_2)}{\mathrm{V}(X_i)},$$

ser vi, at vi har fået generaliseret variansreduktionsegenskaben. Det er iøvrigt vigtigt at bemærke, at det af determinantformlen fremgår, at man kan beregne den multiple korrelationskoefficient ud fra **korrelations**matricen ved de samme formler, som gælder ved beregning ud fra **dispersions**matricen. ▼

Angående **estimation** af multiple korrelationskoefficienter gælder fuldstændig samme bemærkning som anført p. 85 angående partielle koefficienter.

I nedenstående eksempel fortsættes eksempel 2.4.

EKSEMPEL 2.6. For at få et indtryk af, i hvor høj grad mineralindholdene C_3A og C_3S i eksempel 2.4 kan forklare variationer i f.eks. 3-døgnsstyrken kan vi beregne den multiple korrelationskoefficient mellem styrke 3 og (C_3S, og C_3A). Vi finder

$$1 - \hat{\rho}^2_{4|12} = \frac{\det \begin{bmatrix} 1 & 0.158 & 0.120 \\ 0.158 & 1 & -0.309 \\ 0.120 & -0.309 & 1 \end{bmatrix}}{1 \cdot \det \begin{bmatrix} 1 & -0.309 \\ -0.309 & 1 \end{bmatrix}}$$

hvor nummereringen af de variable svarer til den i eksempel 2.3 anvendte. Vi finder

$$\hat{\rho}^2_{4|12} = 1 - 0.9435 = 0.0565.$$

Data tyder derfor på, at kun knap 6% af variationen i cementstyrken (af prøver indsamlet på den måde de konkrete data er indsamlet) kan forklares ved variationer i C_3S og C_3A indholdet alene. ♦

Hvis den multiple korrelationskoefficient er 0, (i.e. hvis $\sigma_i = 0$) er det ikke vanskeligt at bestemme fordelingen af $\hat{\rho}^2_{i|m+1,\ldots,p}$. Vi giver resultatet i en lidt ændret form i

SÆTNING 2.23. Lad $R = \hat{\rho}_{i|m+1,\ldots,p}$ være den empiriske multiple kor relationskoefficient mellem X_i og $\mathbf{X}_2 = (X_{m+1}, \ldots, X_p)$ baseret på n observationer. Da er

$$\frac{R^2}{1-R^2} \cdot \frac{n-(p-m)-1}{p-m} \in \mathrm{F}(p-m, n-(p-m)-1),$$

såfremt $\rho_{i|m+1,\ldots,p} = 0$. ▲

BEVIS 2.20. Forbigås. ■

Dette kan bruges ved test af hypotesen

$$H_0 : \rho_{i|m+1,\ldots,p} = 0 \quad \text{mod} \quad H_1 : \rho_{i|m+1,\ldots,p} \neq 0.$$

Vi forkaster hypotesen for store værdier af teststørrelsen. Vi illustrerer dette i

EKSEMPEL 2.7. Vi betragter situationen fra eksempel 2.6 og vil undersøge, om det kan antages, at den multiple korrelation mellem X_1 og (X_2, X_3) er 0. Vi finder teststørrelsen

$$\frac{R^2}{1-R^2}\frac{51-(3-1)-1}{3-1} = \frac{0.0565}{0.9435}\cdot\frac{48}{2} = 1.44.$$

Da

$$\mathrm{F}(2,48)_{0.90} = 2.42,$$

vil vi i det mindste acceptere en hypotese om, at $\rho_{3|12} = 0$ på alle niveauer $\alpha < 10\%$. Med det foreliggende materiale kan det altså ikke afvises, at $\rho_{3|12} = 0$. Dette udelukker selvsagt ikke, at den er $\neq 0$ (hvad den givetvis er); vi kan blot ikke med sikkerhed konstatere det med de foreliggende data, fordi $\rho_{3|12}$ formentlig er ret lille. ♦

Vi skal ikke komme ind på spørgsmål angående test for andre værdier af $\rho_{i|m+1,\ldots,n}$.

2.3.3 Regression

Vi skal ikke her give nogen særligt dybtgående gennemgang af den såkaldte regressionsteori, der ikke må forveksles med den i følgende afsnit omtalte (lineære) regressionsanalyse.

Lad $\begin{bmatrix} Y \\ \mathbf{X} \end{bmatrix}$ være en stokastisk vektor. Ved **regressionen af Y på $\mathbf{x}$** forstås da funktionen, der er givet ved

$$\mathrm{g}(\mathbf{x}) = \mathrm{E}(Y|\mathbf{X} = \mathbf{x}),$$

d.v.s. **den betingede middelværdi som funktion af den betingede variabel**.

Lad specielt $\begin{bmatrix} Y \\ \mathbf{X} \end{bmatrix}$ være normalt fordelt med parametre

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \text{og} \quad \Sigma = \begin{bmatrix} \sigma_{11} & \sigma_1' \\ \sigma_1 & \Sigma_{22} \end{bmatrix}.$$

Da viser sætning 2.16, at

$$\mathrm{g}(\mathbf{x}) = \mathrm{E}(Y|\mathbf{X} = \mathbf{x}) = \mu_1 + \sigma_1'\Sigma_{22}^{-1}(\mathbf{x} - \mu_2),$$

d.v.s. **regressionen er lineær (affin).**

Vi specialiserer nu til 2 dimensionener.

Lad $\begin{bmatrix} Y \\ X \end{bmatrix}$ være normalt fordelt med parametre

$$\begin{bmatrix} \mu_y \\ \mu_x \end{bmatrix} \quad \text{og} \quad \begin{bmatrix} \sigma_y^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_x^2 \end{bmatrix}.$$

Da er regressionen af Y på X givet ved

$$\mathrm{E}(Y|X=x) = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x-\mu_x),$$

og regressionen af X på Y ved

$$\mathrm{E}(X|Y=y) = \mu_x + \rho\frac{\sigma_x}{\sigma_y}(y-\mu_y).$$

Lad os antage, at der foreligger målinger $\begin{bmatrix} X_1 \\ Y_1 \end{bmatrix}, \ldots, \begin{bmatrix} X_n \\ Y_n \end{bmatrix}$.

Maksimum likelihood estimatorer for hældningskoefficienterne fås da ved blot at indsætte maksimum likelihood estimatorer for de indgående parametre. Da

$$\begin{aligned} \hat{\rho} &= \frac{\sum(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum(X_i-\bar{X})^2\sum(Y_i-\bar{Y})^2}} = \frac{\mathrm{SAP}_{xy}}{\sqrt{\mathrm{SAK}_x\mathrm{SAK}_y}}, \\ \hat{\sigma}_x^2 &= \frac{1}{n}\sum(X_i-\bar{X})^2, \\ \hat{\sigma_y^2} &= \frac{1}{n}\sum(Y_i-\bar{Y})^2, \end{aligned}$$

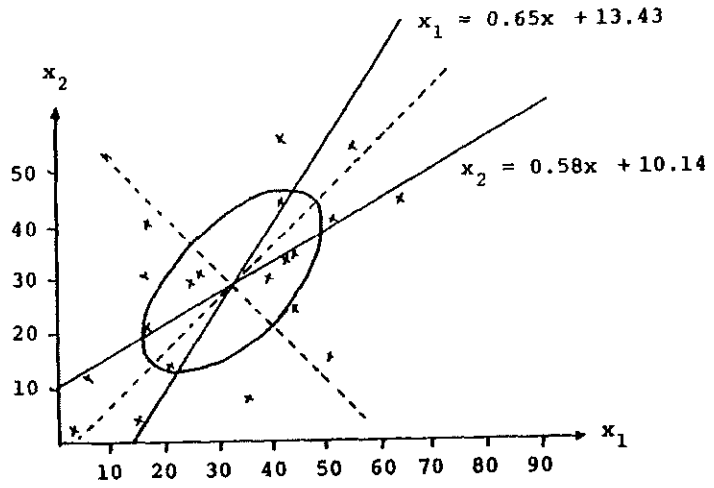
ses, at f.eks. skønnet over hældningskoefficienten i udtrykket for regressionen af Y på X bliver

$$\hat{\rho}\frac{\hat{\sigma_y}}{\hat{\sigma_x}} = \frac{\mathrm{SAP}_{xy}}{\mathrm{SAK}_x}.$$

Dette giver den empiriske regressionsligning

$$\hat{\mathrm{E}}(Y|X=x) = \bar{Y} + \frac{\mathrm{SAP}_{xy}}{\mathrm{SAK}_x}(x-\bar{X}),$$

d.v.s. præcis samme resultat, som vi opnåede i den endimensionale lineære regressionsanalyse, jvf. afsnit 2 i bind 1. Her var forudsætningerne dog ganske anderledes, idet det



Figur 2.5:

blev antaget, at værdierne af den uafhængige variable (her X, i bind 1) var deterministiske værdier, hvor vi i denne fremstilling regner med, at de er realiserede udfald af en normalt fordelt variabel, der er korreleret med den afhængige variabel. Hvad angår estimationen er det altså underordnet, hvilken af de to modeller man arbejder med; men tolkningen af de færdige resultater afhænger selvsagt heraf.

Vi forsætter nu betragtningerne i eksempel 2.8.

EKSEMPEL 2.8. Vi vil i dette eksempel bestemme de lineære overgange fra en måling ved den ene af de to i eksempel 2.2 anførte målemetoder til den anden.

Vi finder regressionerne

$$\begin{aligned}\hat{\mathrm{E}}(X_1|X_2 = x_2) &= \bar{x_1} + \hat{\rho}\frac{s_1}{s_2}(x_2 - \bar{x_2}) \\ &= 0.65x_2 + 13.43\end{aligned}$$

og

$$\begin{aligned}\hat{\mathrm{E}}(X_2|X_1 = x_1) &= \bar{x_2} + \hat{\rho}\frac{s_2}{s_1}(x_1 - \bar{x_1}) \\ &= 0.58x_1 + 10.14.\end{aligned}$$

Disse linier er angivet i figur 2.8. Hvis vi vil undersøge, om der overhovedet er en sammenhæng mellem X_1 og X_2 kan vi undersøge korrelationskoefficienten. Den er

fundet til

$$\hat{\rho} = \frac{182}{\sqrt{311 \cdot 279}} = 0.617,$$

d.v.s.

$$\hat{\rho}^2 = 0.380.$$

Teststørrelsen for test af hypotesen $\rho = 0$ er, jvf. p. 88, med $p = m = 2$

$$t = \frac{0.617}{\sqrt{1 - 0.380}}\sqrt{20 - 2} = 3.32 > \mathrm{t}(18)_{99.5}.$$

Ved test på niveau $\alpha > 1\%$ vil hypotesen derfor blive forkastet, og vi vil antage, at $\rho \neq 0$, d.v.s. at der er en **lineær** sammenhæng mellem måleresultater ved de to metoder, og den er estimeret ved de to regressioner. Man kan så på sædvanlig vis finde skøn over usikkerheder m.v.

I figuren er også indtegnet en konturellipse og dens hovedakser. Det kan vises, at første hovedakse er den linie, man får frem ved at **minimalisere punkternes ortogonale afstandes kvadratsum**, hvor regressionsligningerne jo findes ved at minimalisere de lodrette henholdsvis vandrette afstande. Første hovedakse kaldes derfor også den ortogonale regression. Dette begreb vender vi tilbage til i kapitel 4. ♦

2.4 Spaltningssætningen

I dette afsnit betragter vi en stokastisk variabel $\mathbf{x} \in \mathrm{N}(\mu, \Sigma)$, hvor Σ er regulær af orden n. Vi betragter det indre produkt defineret ved Σ^{-1} og den derudfra affødte norm, nemlig

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

og

$$\|\mathbf{x}\| = \sqrt{(\mathbf{x}|\mathbf{x})} = \sqrt{\mathbf{x}'\Sigma^{-1}\mathbf{x}}$$

Lad nu underrummene $U_1, \ldots, U_k$ være **ortogonale** (m.h.t. dette indre produkt), således at

$$R = U_1 \oplus \ldots \oplus U_k.$$

Vi sætter dim $U_i = n_i$ og betegner projektionen på U_i med p_i. Den tilsvarende matrix benævnes $\mathbf{C}_i$.

Lad betegnelserne være som ovenfor. Da gælder

SÆTNING 2.24. **(Spaltningssætningen)** Sættes

$$\mathbf{Y}_i = p_i(\mathbf{x} - \mu), \qquad i = 1, \ldots, k$$

og

$$K_i = \|\mathbf{Y}_i\|^2 = \|p_i(\mathbf{x} - \mu)\|^2, \qquad i = 1, \ldots, k,$$

da gælder, at

$$\mathbf{x} - \mu = \sum_{i=1}^{k} \mathbf{Y}_i$$

og

$$\|\mathbf{x} - \mu\|^2 = \sum_{i=1}^{k} K_i.$$

Endvidere er $\mathbf{Y}_1, \ldots, \mathbf{Y}_k$ stokastisk uafhængige og normalt fordelte, og $K_1, \ldots, K_k$ er stokastisk uafhængige $\chi^2(n_i)$-fordelte variable. ▲

BEVIS 2.21. Vi har at $\mathbf{Y}_i = \mathbf{C}_i(\mathbf{x} - \mu)$, hvorfor

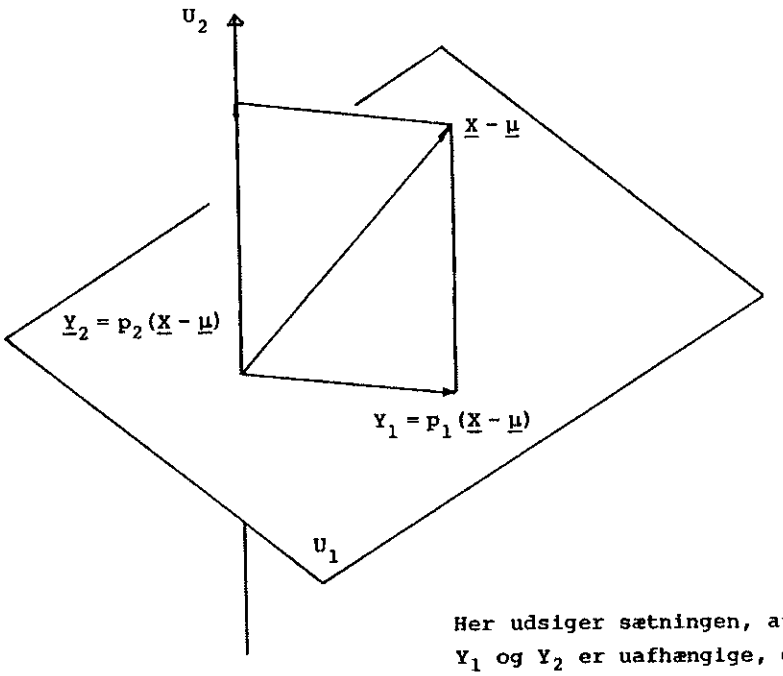
$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ \vdots \\ \mathbf{Y}_k \end{bmatrix} = \begin{bmatrix} \mathbf{C}_1 \\ \vdots \\ \mathbf{C}_k \end{bmatrix} (\mathbf{X} - \mu).$$

heraf fås

$$\mathrm{D}(\mathbf{Y}) = \begin{bmatrix} \mathbf{C}_1 \\ \vdots \\ \mathbf{C}_k \end{bmatrix} \cdot \boldsymbol{\Sigma} \cdot (\mathbf{C}_1', \ldots, \mathbf{C}_k') = (\mathbf{C}_i \boldsymbol{\Sigma} \mathbf{C}_j')_{(i,j)}.$$

Nu er for $i \neq j$ ifølge lemma p. 53.

$$\mathbf{C}_i \boldsymbol{\Sigma} \mathbf{C}_j' = \mathbf{0}.$$



Her udsiger sætningen, at
Y_1 og Y_2 er uafhængige, og
at $\| Y_1 \|^2 \in \chi^2(2)$ og
$\| Y_2 \|^2 \in \chi^2(1)$.

Figur 2.6:

Heraf følger at $\mathbf{Y}$'s komponenter er stokastisk uafhængige (fordi $\mathbf{Y}$ er normalt fordelt).

Vi skal dernæst bestemme fordelingen af $\|p_i(\mathbf{X}-\mu)\|^2$. Vi har, at $\mathbf{X}$ kan skrives

$$\mathbf{X} = \mu + \mathbf{A}\mathbf{Z}$$

hvor $\mathbf{Z} \in \mathrm{N}(0, \mathbf{I})$ og $\mathbf{A}\,\mathbf{A}' = \boldsymbol{\Sigma}$. Heraf følger

$$\begin{aligned} \|p_i(\mathbf{X}-\mu)\|^2 &= \|p_i(\mathbf{A}\mathbf{Z})\|^2 = \|\mathbf{C}_i\mathbf{A}\mathbf{Z}\|^2 \\ &= \mathbf{Z}'\mathbf{A}'\mathbf{C}_i'\boldsymbol{\Sigma}^{-1}\mathbf{C}_i\mathbf{A}\mathbf{Z} = \mathbf{Z}'\mathbf{D}_i\mathbf{Z}. \end{aligned}$$

Nu er

$$\begin{aligned} \mathbf{D}_i\mathbf{D}_i &= \mathbf{A}'\mathbf{C}_i'\boldsymbol{\Sigma}^{-1}\mathbf{C}_i\mathbf{A}\mathbf{A}'\mathbf{C}_i'\boldsymbol{\Sigma}^{-1}\mathbf{C}_i\mathbf{A} \\ &= \mathbf{A}'\mathbf{C}_i'\mathbf{C}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}\,\mathbf{C}_i'\boldsymbol{\Sigma}^{-1}\mathbf{C}_i\,\mathbf{A} \\ &= \mathbf{A}'\mathbf{C}_i'\boldsymbol{\Sigma}^{-1}\mathbf{C}_i\mathbf{A} \\ &= \mathbf{D}_i, \end{aligned}$$

d.v.s. $\mathbf{D}_i$ er idempotent. Vi har i ovenstående gentagne gange benyttet lemma p. 53. Det er åbenbart, at $\mathrm{rg}(\mathbf{D}_i) = n_i$. Da

$$\begin{aligned} \mathbf{D}_i &= \mathbf{A}'\mathbf{C}_i'\mathbf{A}'^{-1}\mathbf{A}^{-1}\mathbf{C}_i\mathbf{A} \\ &= (\mathbf{A}^{-1}\mathbf{C}_i\mathbf{A})'(\mathbf{A}^{-1}\mathbf{C}_i\mathbf{A}), \end{aligned}$$

er $\mathbf{D}_i$ positiv semidefinit (jvf. sætning 1.16 p. 38). Der eksisterer derfor en ortogonal matrix $\mathbf{P}'$ (sætning 1.10) så

$$\mathbf{P}'\mathbf{D}_i\mathbf{P} = \boldsymbol{\Lambda}_i \quad \text{eller} \quad \mathbf{D}_i = \mathbf{P}\,\boldsymbol{\Lambda}_i\mathbf{P}',$$

hvor $\boldsymbol{\Lambda}_i$ er en diagonalmatrix med rang n_i. Da $\mathbf{D}_i$ er idempotent, fås

$$\mathbf{P}\,\boldsymbol{\Lambda}_i\mathbf{P}' = \mathbf{P}\,\boldsymbol{\Lambda}_i\mathbf{P}'\mathbf{P}\,\boldsymbol{\Lambda}_i\mathbf{P}' = \mathbf{P}\,\boldsymbol{\Lambda}_i^2\mathbf{P}',$$

eller $\boldsymbol{\Lambda}_1 = \boldsymbol{\Lambda}_i^2$. Derfor har $\boldsymbol{\Lambda}_i$ n_i 1-taller og $n - n_i$ 0'ler i diagonalen. Derfor er

$$\begin{aligned} \mathbf{Z}'\mathbf{D}_i\mathbf{Z} &= \mathbf{Z}'\mathbf{P}\,\boldsymbol{\Lambda}_i\mathbf{P}'\mathbf{Z} = (\mathbf{P}'\mathbf{Z})'\boldsymbol{\Lambda}_i(\mathbf{P}'\mathbf{Z})' \\ &= \mathbf{V}'\boldsymbol{\Lambda}_i\mathbf{V} \\ &= \underbrace{V_1^2 + \cdots + V_n^2}_{n_i \text{ led} \neq 0.} \end{aligned}$$

Da $\mathbf{V} \in \mathrm{N}(\mathbf{0}, \mathbf{P}'\mathbf{P}) = \mathrm{N}(\mathbf{0}, \mathbf{I})$ ses det, at

$$\mathbf{Z}'\mathbf{D}_i\mathbf{Z} = \|p_i(\mathbf{X}-\mu)\|^2 \in \chi^2(n_i).$$

■

EKSEMPEL 2.9. Lad $X_1, \ldots, X_n$ være uafhængige $\mathrm{N}(\mu, \sigma^2)$-fordelte. Da er

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \in \mathrm{N}(\mu, \sigma^2 \mathbf{I}).$$

Vi betragter underrummet U_1 givet ved

$$\mathbf{x} \in U_i \quad \Leftrightarrow \quad x_i = \ldots = x_n,$$

og det på U_1 vinkelrette (m.h.t. $\sigma^2\mathbf{I}$) U_2. (dette ortogonalitetsbegreb falder sammen med det sædvanlige). Nu viser identiteten

$$\sum (x_i - y)^2 = \sum (x_i - \bar{x})^2 + n(\bar{x} - y)^2,$$

at projektionen på U_1 er givet ved

$$p_1(\mathbf{x}) = \begin{bmatrix} \bar{x} \\ \vdots \\ \bar{x} \end{bmatrix},$$

hvorfor

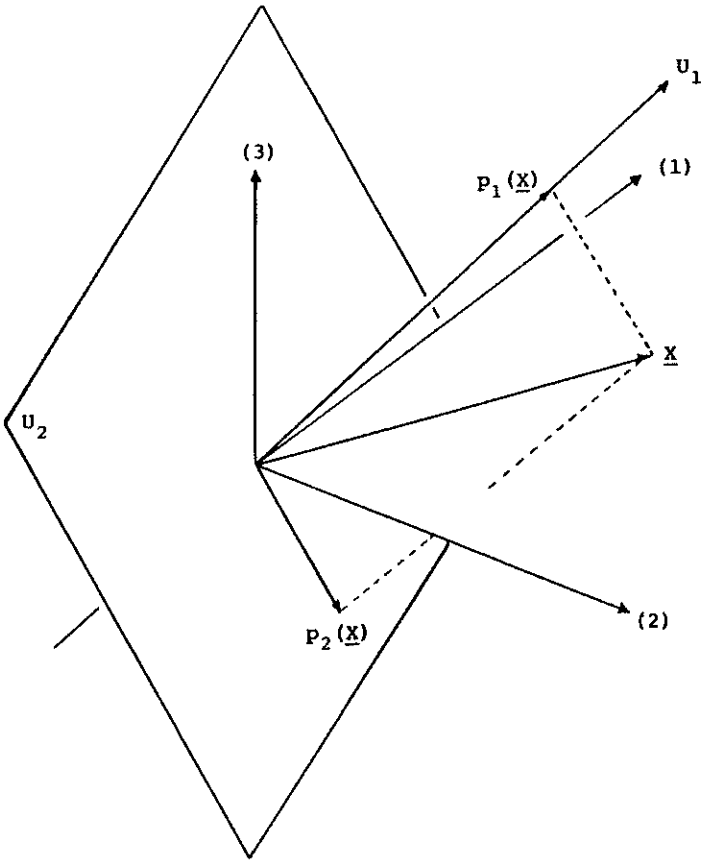
$$p_2(\mathbf{x}) = \mathbf{x} - p_i(\mathbf{x}) = \begin{bmatrix} x_1 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{bmatrix}.$$

Da $\dim U_1 = 1$ og $\dim U_2 = n - 1$ får vi af spaltningssætningen, at

$$p_1(\mathbf{X} - \mu) \quad \text{og} \quad \|p_2(\mathbf{X} - \mu)\|^2$$

er stokastisk uafhængige. $p_1(\mathbf{X} - \mu)$ er normalt fordelt og $\|p_2(\mathbf{X} - \mu)\|^2$ er $\chi^2(n-1)$-fordelt. Da

$$p_1(\mathbf{X} - \mu) = \begin{bmatrix} \bar{X} - \mu \\ \vdots \\ \bar{X} - \mu \end{bmatrix},$$



Figur 2.7:

og

$$\|p_2(\mathbf{X}-\mu)\|^2 = \frac{1}{\sigma^2}\sum_1(X_i-\bar{X})^2,$$

genfinder vi her et resultat om fordelingen af $\bar{X}$ og $(n-1)S^2 = \frac{1}{\sigma^2}\sum(X_i-\bar{X})^2$. ♦

2.5 Wishart fordelingen og den generaliserede varians

I det endimensionale tilfælde udledes en række stikprøvefordelinger fra det normale tilfælde. Den vigtigste af disse er χ^2-fordelingen, der svarer til summen af kvadrerede normalt fordelte variable. Den flerdimensionale analog er Wishart fordelingen.

Vi giver definitionen ved hjælp af tætheden i

DEFINITION 2.3. Lad $\mathbf{V}$ være en kontinuert fordelt stokastisk $p \times p$-matrix, der er symmetrisk og positivt semidefinit med sandsynlighed 1. Da siges $\mathbf{V}$ at følge en **Wishart fordeling** med parametre (n, Σ), $(n \geq p)$, hvis tætheden for $\mathbf{V}$ er

$$f(\mathbf{v}) = c \cdot [\det(\mathbf{v})]^{\frac{1}{2}(n-p-1)} \exp(-\frac{1}{2}\operatorname{tr}(\mathbf{v} \cdot \Sigma^{-1})),$$

for $\mathbf{v}$ positivt definit og 0 ellers. Her betegner Σ en positivt definit $p \times p$-matrix, og c er konstanten givet ved

$$\frac{1}{c} = 2^{\frac{1}{2}np}\pi^{p(p-1)/4}(\det \Sigma)^{\frac{1}{2}n} \prod_{i=1}^{p} \Gamma(\frac{1}{2}(n+1-i)).$$

Vi skriver kort

$$\mathbf{V} \in \mathrm{W}(n, \Sigma) = \mathrm{W}_p(n, \Sigma).$$

hvor den sidste betegnelse anvendes, hvis der kan opstå tvivl om dimensionen.

Vi anfører nu først en bemærkning om middelværdi og varians af komponenterne i en Wishart fordeling

Lad $\mathbf{V} = (V_{ij})$ være Wishart fordelt $\mathrm{W}(n, \Sigma)$, hvor $\Sigma = (\sigma_{ij})$. Da gælder

$$\begin{aligned} \mathrm{E}(V_{ij}) &= n\sigma_{ij} \\ \mathrm{V}(V_{ij}) &= n(\sigma_{ij}^2 + \sigma_{ii}\sigma_{jj}) \\ \mathrm{Cov}(V_{ij}, V_{kl}) &= n(\sigma_{ik}\sigma_{jl} + \sigma_{il}\sigma_{jk}). \end{aligned}$$

▲

BEVIS 2.22. Forbigås. ■

Analogien med χ^2-fordelingen fremgår af

SÆTNING 2.25. Lad $\mathbf{X}_i \in \mathrm{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$, $i = 1, \ldots, n$, være uafhængige og regulært fordelte. Da gælder for $n \geq p$

$$\mathbf{Y} = \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i' \in \mathrm{W}(n, \boldsymbol{\Sigma}).$$

▲

BEVIS 2.23. Forbigås. ■

BEMÆRKNING 2.12. Hvis $n \leq p$ har $\mathbf{Y}$ som defineret i sætningen ingen tæthedsfunktion. Vi vælger dog alligevel at sige, at $\mathbf{Y}$ er Wishart fordelt med parametre $(n, \boldsymbol{\Sigma})$.

Tilsvarende bemærkninger gør sig gældende hvis $\boldsymbol{\Sigma}$ er singulær. Med denne vedtagelse gælder sætningen uden restriktionen $n \leq p$. ▼

En næsten triviel følge af ovenstående er nu

SÆTNING 2.26. Lad $\mathbf{V}_1, \ldots, \mathbf{V}_k$ være uafhængige stokastiske $p \times p$-matricer, der er $\mathrm{W}(n_i, \boldsymbol{\Sigma})$-fordelte. Da gælder

$$\mathbf{V} = \mathbf{V}_1 + \cdots + \mathbf{V}_k \in \mathrm{W}(n_1 + \cdots + n_k, \boldsymbol{\Sigma}).$$

En af hovedsætningerne i teorien for stikprøvefunktioner af normalt fordelte stokastiske variable er, at $\bar{X}$ og S^2 er uafhængige og at S^2 er $\sigma^2\chi^2/f$-fordelt med 1 frihedsgrad mindre end der er observationer. Denne sætning har sin flerdimensionale analog i ▲

SÆTNING 2.27. Lad $\mathbf{X}_i \in \mathrm{N}_p(\boldsymbol{\mu}, \Sigma)$, $i = 1, \ldots, n$, være stokastisk uafhængige. Vi sætter

$$\begin{aligned} \bar{\mathbf{X}} &= \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i, \\ \mathbf{S} &= \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})'. \end{aligned}$$

Da er

$$\bar{\mathbf{x}} \in \mathrm{N}_p(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$$

og

$$\mathbf{S} \in \mathrm{W}(n-1, \frac{1}{n-1}\Sigma).$$

Endvidere er $\bar{\mathbf{X}}$ og $\mathbf{S}$ stokastisk uafhængige. ▲

BEVIS 2.24. Forbigås. ■

Vi går nu over til at betragte nogle resultater vedrørende marginale fordelinger. Vi har

SÆTNING 2.28. Lad $\mathbf{V}$ være Wishart fordelt med parametre (n, Σ). Vi betragter spaltningen

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} \\ \mathbf{V}_{21} & \mathbf{V}_{22} \end{bmatrix} \quad \text{og} \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

Det gælder da, at

$$\mathbf{V}_{ii} \in \mathrm{W}(n, \Sigma_{ii}).$$

▲

Endvidere gælder

SÆTNING 2.29. Vi betragter igen ovenstående situation. Hvis Σ_{12} og Σ_{21} er $\mathbf{0}$-matricer, da er $\mathbf{V}_{11}$ og $\mathbf{V}_{22}$ stokastisk uafhængige. ▲

BEVIS 2.25. for sætningerne. De følger umiddelbart ved at betragte de tilsvarende spaltninger af normalt fordelte vektorer, som frembringer Wishart fordelingerne. ■

Da den flerdimensionale normalfordeling kan defineres på en koordinatfri måde, i.e. uafhængigt af et givent koordinatsystem, er det ikke overraskende, at noget tilsvarende gør sig gældende for Wishart fordelingen. Da skift fra koordinater i et koordinatsystem til koordinater i et andet foregår hved hjælp af matrixmultiplikationer, har vi følgende

SÆTNING 2.30. Lad $\mathbf{V} \in \mathrm{W}_p(n, \Sigma)$ og lad $\mathbf{A}$ være en vilkårlig, fast $r \times p$-matrix. Da er

$$\mathbf{A}\,\mathbf{V}\,\mathbf{A}' \in \mathrm{W}_r(n, \mathbf{A}\,\Sigma\,\mathbf{A}').$$

▲

BEVIS 2.26. Som antydet ovenfor skal man blot betragte de normalt fordelte vektorer, der frembringer V og transformere dem. Resultatet følger da umiddelbart. ■

Vi afslutter nu kapitlet med at indføre en anden generalisering af den endimensionale varians til det flerdimensionale tilfælde end dispersionsmatricen.

DEFINITION 2.4. Lad den p-dimensionale vektor $\mathbf{X}$ have dispersionsmatricen $\boldsymbol{\Sigma}$. Ved **den generaliserede varians** af $\mathbf{X}$ forstås determinanten af dispersionsmatricen, i.e.

$$\text{gen.var.}(\mathbf{X}) = \det(\boldsymbol{\Sigma}).$$

▲

BEMÆRKNING 2.13. I afsnit 1.2.6 godtgjordes, at determinanten af en matrix svarer til volumenforholdet ved den tilsvarende lineære afbildning, d.v.s. den er et intuitivt rimeligt mål for "størrelsen" af en matrix. ▼

Hvis der foreligger observationer $\mathbf{X}_1, \ldots, \mathbf{X}_n$, defineres den **empiriske generaliserede varians** på åbenbar måde ud fra den empiriske dispersionsmatrix

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

nemlig som dennes determinant.

I det normale tilfælde kan vi finde fordelingen af den empiriske generaliserede varians, idet vi har

SÆTNING 2.31. Lad $\mathbf{X}_i \in \mathrm{N}_p(\mu, \boldsymbol{\Sigma})$, $i = 1, \ldots, n$, være stokastisk uafhængige. Da følger den empiriske generaliserede varians samme fordeling som

$$\frac{\det \boldsymbol{\Sigma}}{(n-1)p} \cdot Z_1 \ldots Z_p,$$

hvor $Z_1, \ldots, Z_p$ er stokastisk uafhængige og $Z_i \in \chi^2(n-i)$. ▲

BEVIS 2.27. Forbigås. ■

For $p = 1$ og 2 er det muligt at finde tætheden af den empiriske generaliserede varians. For større værdier af p involverer denne tæthed dog integraler, som det ikke umiddelbart er muligt at udtrykke ved kendte funktioner. For $n \to \infty$ har vi dog

SÆTNING 2.32. Lad S være som ovenfor (i det normale tilfælde). Da gælder, at

$$\sqrt{n-1}\left(\frac{\det(\mathrm{S})}{\det(\Sigma)} - 1\right) \quad \text{asymptotisk} \quad \in \mathrm{N}(0, 2p).$$

▲

BEVIS 2.28. Forbigås. ■

2.6 Lidt om estimation af flerdimensionale parametre

I dette afsnit skal vi anføre udvidelser til tilfældet med flerdimensionale parametre af de i bind 1, kapitel 2, anførte resultater om nedre grænser for variansen af en central estimator (Cramér-Rao's ulighed) og af sætningen om den asymptotiske normalitet af maximum likelihood estimatorer.

Sætningerne synes teknisk meget komplicerede, men kvintessen af de regularitetsbetingelser, der anføres, er, at differentiation med hensyn til parametrene kan ombyttes med integration efter x'erne.

Vi må først indføre begrebet informationsmatricen for en flerdimensional parameter.

DEFINITION 2.5. Lad $\mathbf{X}_1, \ldots, \mathbf{X}_n$ være indbyrdes uafhængige, identisk fordelte stokastiske variable med frekvensfunktion $\mathrm{f}(\mathbf{x}, \theta)$, hvor $\theta \in \Omega \subseteq R^k$. Ved **informationsmatricen** $\mathbf{i}(\theta)$ forstås da matricen med (i, j)'te element

$$\mathbf{i}(\theta)_{i,j} = \mathrm{E}_\theta \left\{ \frac{\partial \log \mathrm{f}(\mathbf{X}_1, \theta)}{\partial \theta_i} \frac{\partial \log \mathrm{f}(\mathbf{X}_1, \theta)}{\partial \theta_j} \right\}.$$

▲

LEMMA 2.1. Under passende regularitetsbetingelser, f.eks. de i efterfølgende sætninger anførte, gælder

$$\mathbf{i}(\theta)_{i,j} = -\mathrm{E}_\theta \left\{ \frac{\partial^2 \log \mathrm{f}(\mathbf{X}_1, \theta)}{\partial \theta_i \, \partial \theta_j} \right\}.$$

BEVIS 2.29. Forbigås. ■

SÆTNING 2.33. Cramér-Rao's ulighed. Lad $\mathbf{X}_1, \ldots, \mathbf{X}_n$ være indbyrdes uafhængige, identisk fordelte stokastiske variable med frekvensfunktion $\mathrm{f}(\mathrm{x}, \theta)$, hvor θ er en ukendt parameter, $\theta \in \Omega \subseteq R^k$. Lad endvidere for alle i, j og for alle $\theta \in \overset{o}{\Omega}$ (det indre af Ω)

1) $\dfrac{\partial \mathrm{f}(\mathrm{x}, \theta)}{\partial \theta_i}$ eksistere for alle $\mathbf{x}$,

2) $\mathrm{E}_\theta \left\{ \dfrac{\partial \log \mathrm{f}(\mathbf{X}_1, \theta)}{\partial \theta_i} \right\} = 0,$

3) $\mathrm{E}_\theta \left\{ \left(\dfrac{\partial \log \mathrm{f}(\mathbf{X}_1, \theta)}{\partial \theta_i} \right)^2 \right\} < \infty,$

4) $\det \mathbf{i}(\theta) \neq 0.$

Lad endvidere $\mathbf{d}(\mathbf{X}_1, \ldots, \mathbf{X}_n) = (d_1, \ldots, d_k)'$ være et centralt skøn over θ, der tilfredsstiller

5) $\mathrm{E}_\theta \left\{ d_i \cdot \dfrac{\partial}{\partial \theta_j} \log \prod_\nu \mathrm{f}(\mathbf{X}_\nu, \theta) \right\} = \delta_{ij},$

6) $\mathrm{E}_\theta \{ d_i^2 \} < \infty.$

Da er

$$\mathrm{D}(\mathbf{d}(\mathbf{X}_1, \ldots, \mathbf{X}_n)) \geq \frac{1}{n} \mathbf{i}^{-1}(\theta),$$

d.v.s. matricen

$$\mathrm{D}(\mathbf{d}(\mathbf{X}_1, \ldots, \mathbf{X}_n)) - \frac{1}{n} \mathbf{i}^{-1}(\theta)$$

er positiv semidefinit. ▲

BEVIS 2.30. Forbigås, se f.eks. [35]. ■

BEMÆRKNING 2.14. Som allerede anført indledningsvis, er de anførte betingelser ikke synderligt restriktive. Hvis man antager, at fordelingen f har en støtte, der er uafhængig af θ, sikrer 1)–3) og 5)–6) blot, at differentiation med hensyn til θ og integration med hensyn til $\mathbf{x}$ kan ombyttes. Eksempelvis har vi i det kontinuerte tilfælde med $\Pi \mathrm{f}(\mathbf{X}_\nu;\theta) = \mathrm{p}(\mathbf{X}_1,\ldots,\mathbf{X}_n;\theta) = P$

$$\begin{aligned} E_\theta\left\{d_i\frac{\partial \log p}{\partial\theta_j}\right\} &= \int d_i\frac{\partial \log p}{\partial\theta_j}p\mathrm{d}\mathbf{z} \\ &= \int d_i\frac{1}{p}\frac{\partial p}{\partial\theta_j}p\mathrm{d}\mathbf{z} \\ &= \frac{\partial}{\partial\theta_j}\int d_i p\mathrm{d}\mathbf{z} \\ &= \frac{\partial}{\partial\theta_j}\theta_i \\ &= \delta_{ij}, \end{aligned}$$

hvor δ_{ij} altså er Kroneckers δ ($=0$ for $i\neq j$, og $=1$ for $i=j$). Integrationen foregår med hensyn til $\mathbf{z}$ givet ved $\mathbf{z}' = (\mathbf{x}_1',\ldots,\mathbf{x}_n')$. ▼

BEMÆRKNING 2.15. Ud fra sætningen kan også formuleres resultater om en nedre grænse for den generaliserede varians af $\mathbf{d}$ etc.

Vi anfører dernæst resultatet om den asymptotiske fordeling af maximum likelihood estimatorer. ▼

SÆTNING 2.34. Lad $\mathbf{X}_1,\ldots,\mathbf{X}_n$ være indbyrdes uafhængige, identisk fordelte stokastiske variable med frekvensfunktion $\mathrm{f}(\mathbf{x},\theta)$, $\theta\in\Omega\subseteq R^k$. Lad endvidere (for $\theta\in\overset{o}{\Omega}$ og for alle i,j)

1) $\theta_1\neq\theta_2 \Rightarrow \mathrm{f}(\cdot,\theta_1)\neq\mathrm{f}(\cdot,\theta_2)$,

2) $\dfrac{\partial^2\mathrm{f}(\mathbf{x},\theta)}{\partial\theta_i\partial\theta_j}$ eksistere og være kontinuert,

3) $\mathrm{E}_\theta\left\{\dfrac{1}{\mathrm{f}(\mathbf{X}_1,\theta)}\dfrac{\partial\mathrm{f}(\mathbf{X}_1,\theta)}{\partial\theta_i}\right\}=0$,

4) $\mathrm{E}_\theta\left\{\dfrac{1}{\mathrm{f}(\mathbf{X}_1,\theta)}\dfrac{\partial^2\mathrm{f}(\mathbf{X}_1,\theta)}{\partial\theta_i,\partial\theta_j}\right\}=0$,

5) eksistere en omegn U omkring θ og en funktion $\mathrm{M}(\mathbf{x},\theta)$ med $\mathrm{E}_\theta\{\mathrm{M}(\mathbf{X}_1,\theta)\}<\infty$, så

$$\left|\frac{\partial^2\log\mathrm{f}(\mathbf{x},\theta)}{\partial\theta_i\partial\theta_j}\right|_{\theta=\theta^*}\leq\mathrm{M}(\mathbf{x},\theta),\qquad\forall\theta^*\in U,$$

6) $\det \mathbf{i}(\theta) \neq 0$.

Kaldes θ's maximum likelihood estimator $\hat{\theta}_n$, og er denne konsistent, da vil $\sqrt{n}(\hat{\theta}_n - \theta)$ være asymptotisk normalt fordelt med parametre $(\mathbf{0}, \mathbf{i}^{-1}(\theta))$. ▲

BEVIS 2.31. Forbigås, se f.eks. [35]. ■

BEMÆRKNING 2.16. Der kan anføres betragtninger om betingelserne 1)–6), som er ganske analoge til de efter sætning 2.6 anførte. ▼

BEMÆRKNING 2.17. Det kræves i sætningen, at $\hat{\theta}_n$ er konsistent, d.v.s. konvergerer i sandsynlighed mod den sande parameterværdi $\mathbf{r}$. En sådan betingelse er ofte ganske let at eftervise direkte. Man kan selvfølgelig også anføre yderligere strammede regularitetsbetingelser, som automatisk vil sikre denne konsistens. De svarer helt til betingelsen 5), men vedrører blot de tredie afledede i stedet for de anden, cf. [8]. ▼

BEMÆRKNING 2.18. Sætningen bevarer også sin gyldighed med forholdsvis simple modifikationer (bl.a. af parametrene) i tilfældet med uafhængige, men ikke nødvendigvis identisk fordelte variable $X_1, \ldots, X_n$ ▼

Kapitel 3
Den generelle lineære model

I dette kapitel skal vi formulere en model, som naturligt generaliserer de fra bind 1 kendte varians- og regressionsanalyser. Sætninger og definitioner vil i vidt omfang blive tolket geometrisk for at give en mere intuitiv forståelse af sammenhængene.

3.1 Estimation i den generelle lineære model

Vi giver først en beskrivelse af modellen i

3.1.1 Modelformulering

Vi betragter en n-dimensional stokastisk variabel $\mathbf{Y} \in \mathrm{N}(\mu, \sigma^2\Sigma)$, hvor Σ er kendt. Vi betragter normen givet ved Σ^{-1}, i.e.

$$\|\mathbf{x}\|^2 = \mathbf{x}'\Sigma^{-1}\mathbf{x}.$$

Den ved den inverse dispersionsmatrix $(\sigma^2\Sigma)^{-1}$ definerede norm er givet ved

$$\|\mathbf{x}\|^2_{\sigma^2} = \frac{1}{\sigma^2}\mathbf{x}'\Sigma^{-1}\mathbf{x} = \frac{1}{\sigma^2}\|\mathbf{x}\|^2.$$

De to normer er altså proportionale, og de giver derfor anledning til samme ortogonalitetsbegreb.

Vi vil nu betragte en række problemer i forbindelse med estimation og testning af middelværdien μ i tilfælde, hvor μ er en kendt lineær funktion af ukendte parametre,

i.e.

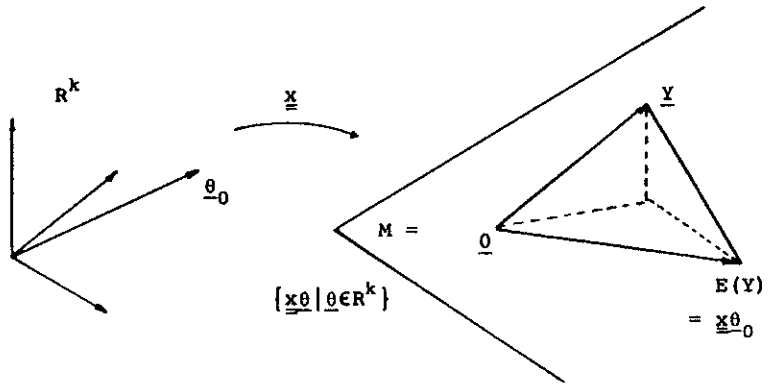
$$\mu = \mathbf{x}\theta$$

eller

$$\begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_k \end{bmatrix},$$

hvor $\mathbf{x}$ altså er kendt.

Geometrisk kan dette udtrykkes, at vi antager, at forventningsværdien af den stokastiske vektor $\mathbf{Y}$ ligger i et underrum M af R^n. M er billedet af R^k ved den til $\mathbf{x}$ svarende lineære afbildning. Dimensionen af M er $\mathrm{rg}(\mathbf{x}) \leq k$. Situationen er anskueliggjort i nedenstående figur.



Figur 3.1: Geometrisk skitse af generel lineær model.

Vi kalder en sådan model, hvor den ukendte middelværdi μ er en (kendt) **lineær** funktion af den ukendte parameter θ for en **(generel) lineær model**. Dette gælder også uden forudsætningen om at $\mathbf{Y}$ er normalt fordelt.

EKSEMPEL 3.1. Vi betragter en sædvanlig endimensional regressionsanalysemodel, d.v.s. der foreligger observationer

$$Y_i = \alpha + \beta x_i + \varepsilon_i, \qquad i = 1, \ldots, n,$$

hvor $\mathrm{E}(\varepsilon_i) = 0$. Denne model kan skrives

$$\begin{bmatrix} Y_i \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

eller

$$\mathbf{Y} = \mathbf{x}\theta + \varepsilon,$$

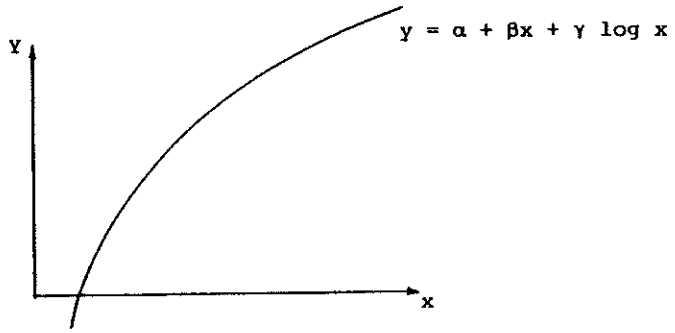
d.v.s. modellen er lineær i ovenstående betydning. ♦

Et andet eksempel er

EKSEMPEL 3.2. Vi betragter nu en situation, hvor

$$Y_i = \alpha + \beta x_i + \gamma \log x_i + \varepsilon_i, \qquad i = 1, \ldots, n$$

stadig med $\mathrm{E}(\varepsilon_i) = 0$. Også i denne situation får vi en **lineær** model frem, nemlig



$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & \log x_1 \\ \vdots & \vdots & \vdots \\ 1 & x_n & \log x_n \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Vi bemærker, at vendingen lineær **ikke** går på, at $\mathrm{E}(Y|X) = \alpha + \beta\, x + \gamma \log\, x$ skal være lineær i den uafhængige variabel x, **men** at $\mathrm{E}(Y|x)$ opfattet som funktion af den ukendte parameter $(\alpha, \beta, \gamma)'$ skal være lineær. Havde vi haft en model som

$$Y_i = \alpha + \beta \log(\gamma x_i + \delta) + \varepsilon_i,$$

hvor α, β, γ og δ er de ukendte parametre, da ville det ikke være muligt at skrive

$$\mathbf{Y} = \mathbf{x} \begin{bmatrix} \alpha \\ \beta \\ \gamma \\ \delta \end{bmatrix} + \varepsilon$$

med en kendt $\mathbf{x}$-matrix, og der vil derfor ikke være tale om en lineær model. ♦

3.1.2 Estimation i det regulære tilfælde

Vi formulerer først resultatet vedrørende estimation af θ i

SÆTNING 3.1. Lad $\mathbf{x}$ og θ være som i foregående afsnit, og lad $\mathbf{Y} \in \mathrm{N}_n(\mathbf{x}\theta, \sigma^2\Sigma)$, hvor Σ er positivt definit. Da er maksimum likelihood estimatoren $\hat{\theta}$ for θ givet ved, at $\mathbf{x}\hat{\theta}$ er projektionen (m.h.t. Σ) ned på M, d.v.s. $\hat{\theta}$ er løsning til den såkaldte **normalligning**

$$(\mathbf{x}'\Sigma^{-1}\mathbf{x})\hat{\theta} = \mathbf{x}'\Sigma^{-1}\mathbf{y}.$$

Hvis $\mathbf{x}$ har fuld rang k, er

$$\hat{\theta} = (\mathbf{x}'\Sigma^{-1}\mathbf{x})^{-1}\mathbf{x}'\Sigma^{-1}\mathbf{Y},$$

og som linearkombination af normalt fordelt variable er $\hat{\theta}$ selv normalt fordelt med parametre

$$\begin{aligned} \mathrm{E}(\hat{\theta}) &= \theta \\ \mathrm{D}(\hat{\theta}) &= \sigma^2(\mathbf{x}'\Sigma^{-1}\mathbf{x})^{-1}. \end{aligned}$$

Det bemærkes specielt, at $\hat{\theta}$ er central for θ. ▲

BEVIS 3.1. Hvis $\mathbf{Y} \in \mathrm{N}(\mathbf{x}\theta, \sigma^2\Sigma)$, hvor Σ er regulær, er tætheden for $\mathbf{Y}$

$$\begin{aligned} \mathrm{f}(\mathbf{y}) &= \frac{1}{\sqrt{2\pi}^n}\frac{1}{\sigma^n}\frac{1}{\sqrt{\det\Sigma}}\exp[-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{x}\theta)'\Sigma^{-1}(\mathbf{y}-\mathbf{x}\theta)] \\ &= k\cdot\frac{1}{\sigma^n}\exp[-\frac{1}{2\sigma^2}\|\mathbf{y}-\mathbf{x}\theta\|^2]. \end{aligned}$$

Vi har likelihood-funktionen

$$\mathrm{L}(\theta) = k\cdot\frac{1}{\sigma^n}\exp[-\frac{1}{2\sigma^2}\|\mathbf{y}-\mathbf{x}\theta\|^2],$$

d.v.s.

$$\log \mathrm{L}(\theta) = k_1 - \frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{x}\theta\|^2.$$

Det er nu åbenbart, at maksimalisering af likelihoodfunktionen er ensbetydende med minimalisring af den kvadratiske afstand mellem et vilkårligt punkt i M og observationen, i.e. ensbetydende med minimalisering af

$$\|\mathbf{y} - \mathbf{x}\theta\|^2 = (\mathbf{y} - \mathbf{x}\theta)'\Sigma^{-1}(\mathbf{y} - \mathbf{x}\theta).$$

Ifølge reslutatet p. 51 er den værdi af $\mathbf{x}\theta$, som giver minimum, lig den ortogonale (m.h.t. Σ^{-1}) projektionen af $\mathbf{y}$ ned på M. Ifølge eksempel 1.8 p. 48 er det optimale θ løsning til ligningen

$$(\mathbf{x}'\Sigma^{-1}\mathbf{x})\theta = \mathbf{x}'\Sigma^{-1}\mathbf{y}.$$

Hvis $\mathbf{x}'\Sigma^{-1}\mathbf{x}$ har fuld rang k, i.e. hvis $\mathbf{x}$ har rang k (jvf. p. 35) er derfor

$$\theta_{\text{opt.}} = (\mathbf{x}'\Sigma^{-1}\mathbf{x})^{-1}\mathbf{x}'\Sigma^{-1}\mathbf{y}.$$

Vi har nu vist den første halvdel af sætningen.

Af sætning 2.2 fås, at

$$\mathrm{E}(\hat{\theta}) = (\mathbf{x}'\Sigma^{-1}\mathbf{x})^{-1}\mathbf{x}'\Sigma^{-1}\mathbf{x}\theta = \theta,$$

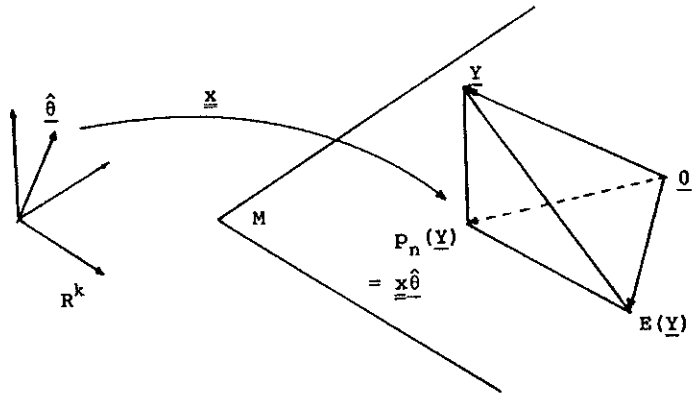
og ved hjælp af sætning 2.5 fås

$$\begin{aligned}\mathrm{D}(\hat{\theta}) &= (\mathbf{x}'\Sigma^{-1}\mathbf{x})^{-1}\mathbf{x}'\Sigma^{-1}(\sigma^2\Sigma)\Sigma^{-1}\mathbf{x}(\mathbf{x}'\Sigma^{-1}\mathbf{x})^{-1} \\ &= \sigma^2(\mathbf{x}'\Sigma^{-1}\mathbf{x})^{-1},\end{aligned}$$

■

Vi har illustreret problematikken i nedenstående fig. 3.2

BEMÆRKNING 3.1. Vi bemærker, at θ er estimeret ved at minimalisere den kvadrerede afstand ned til M. $\hat{\theta}$ er derfor også et **mindste kvadraters** skøn over θ. Hvis man ikke opretholder fordelingsforudsætningen vil man derfor alligevel ofte anvende den i sætning 3.1 anførte estimator $\hat{\theta}$ som skøn over θ. Det kan iøvrigt vises, at mindste kvadraters skønnet $\hat{\theta}$ har den mindste generaliserede varians blandt alle de estimatorer, der er lineære funktioner af observationerne (den såkaldte **Gauss-Markov sætning**), se f.eks. [20]. ▼



Figur 3.2: Geometrisk sktise af estimationsproblematikken i den generelle lineære model.

Da σ^2 oftest er ukendt, vil vi nu udlede estimatorer for denne. Vi har

SÆTNING 3.2. Lad situationen være som ovenfor. Maximum likelihood skønnet for σ^2 er

$$\frac{1}{n}\|\mathbf{Y} - \mathbf{x}\hat{\theta}\|^2.$$

Det centrale skøn over σ^2 er

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n - \operatorname{rg}\mathbf{x}}\|\mathbf{Y} - \mathbf{x}\hat{\theta}\|^2 \\ &= \frac{1}{n - \operatorname{rg}\mathbf{x}}(\mathbf{Y} - \mathbf{x}\hat{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{x}\hat{\theta})\end{aligned}$$

hvor $\mathbf{x}\hat{\theta}$ er maximum likelihood skønnet over $\mathrm{E}(\mathbf{Y})$. Der gælder, at

$$\hat{\sigma}^2 \in \sigma^2\chi^2(n - \operatorname{rg}\mathbf{x})/(n - \operatorname{rg}\mathbf{x})$$

og $\hat{\sigma}^2$ er uafhængig af maximum likelihood skønnet over middelværdien og dermed uafhængig af $\hat{\theta}$. ▲

BEVIS 3.2. Likelihoodfunktionen bliver

$$\mathrm{L}(\theta, \sigma^2) = k \cdot \frac{1}{\sigma^n}\exp[-\frac{1}{2}\frac{1}{\sigma^2}\|\mathbf{y} - \mathbf{x}\theta\|^2],$$

og

$$\log \mathrm{L}(\theta, \sigma^2) = k_1 - \frac{n}{2}\log\sigma^2 - \frac{1}{2}\frac{1}{\sigma^2}\|\mathbf{y} - \mathbf{x}\theta\|^2.$$

Nu er

$$\begin{aligned}\frac{\partial}{\partial\sigma^2}\log \mathrm{L} &= -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\|\mathbf{y} - \mathbf{x}\theta\|^2 \\ &= -\frac{n}{2}\frac{1}{\sigma^4}(\sigma^2 - \frac{1}{n}\|\mathbf{y} - \mathbf{x}\theta\|^2).\end{aligned}$$

Ved differentiation med hensyn til θ fås blot det sædvanlige normalligningssystem. Vi har derfor, at maximum likelihood skønnene $(\hat{\theta}, \hat{\sigma}^2)$ for (θ, σ^2) er løsningen til

$$\begin{aligned}&\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x}\,\hat{\theta} = \mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{Y} \\ &\hat{\sigma}^2 = \frac{1}{n}\|\mathbf{Y} - \mathbf{x}\hat{\theta}\|^2 = \frac{1}{n}(\mathbf{Y} - \mathbf{x}\hat{\theta})\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{x}\hat{\theta}).\end{aligned}$$

Betragter vi opspaltningen af R^n som direkte sum af M og $M^\perp$, hvor $M^\perp$ er det ortogonale (m.h.t. $\boldsymbol{\Sigma}^{-1}$) komplement til M, fås, at

$$\mathrm{P}_M(\mathbf{Y} - \mathbf{x}\theta) = \mathbf{x}\hat{\theta} - \mathbf{x}\theta$$

og

$$\mathbf{Y} - \mathbf{x}\hat{\theta}$$

er stokastisk uafhængige, og at

$$\begin{aligned}&\|\mathbf{Y} - x\hat{\theta}\|^2 \in \sigma^2\chi^2(\dim M^\perp) \\ &= \sigma^2\chi^2(n - \operatorname{rg}\mathbf{x}).\end{aligned}$$

Heraf fås specielt

$$\mathrm{E}(\hat{\sigma}^2) = \frac{1}{n}(n - \operatorname{rg}\mathbf{x})\sigma^2,$$

d.v.s. maximum likelihood estimatoren for σ^2 er ikke central. Ønsker vi et centralt skøn, kan vi anvende

$$\frac{1}{n - \operatorname{rg}\mathbf{x}}\|\mathbf{Y} - \mathbf{x}\hat{\theta}\|^2.$$

Da vi som oftest vil bruge det centrale skøn over σ^2, vil vi knytte betegnelsen $\hat{\sigma}^2$ til dette. ■

BEMÆRKNING 3.2. Hvis Σ specielt er enhedsmatricen, bliver $\|\mathbf{y}\|^2 = \sum y_i^2$. Derfor fås i dette tilfælde

$$\hat{\sigma}^2 = \frac{1}{n - \operatorname{rg} \mathbf{x}} \sum_{i=1}^{n} (Y_i - \hat{\mathrm{E}}(Y_i))^2,$$

hvor $\hat{\mathrm{E}}(Y_i) = (\mathbf{x}\hat{\theta})_i$. Størrelsen $Y_i - \hat{\mathrm{E}}(i)$ er lig den i'te observations afvigelse fra den estimerede model, og den kaldes det i'te **residual**. Vi har altså i tilfældet $\Sigma = \mathbf{I}$, at variansskønnet er proportionalt med **summen af de kvadrerede residualer**, som benævnes $\mathrm{SAK}_{\mathrm{res}}$. Vi vedtager, at vi vil anvende denne betegnelse generelt for den kvadrerede afstand mellem observationen og den estimerede model, d.v.s.

$$\mathrm{SAK}_{\mathrm{res}} = \|\mathbf{Y} - \mathbf{x}\theta\|^2 = (\mathbf{Y} - \mathbf{x}\hat{\theta})'\Sigma^{-1}(\mathbf{Y} - \mathbf{x}\hat{\theta}).$$

▼

Inden vi går videre, giver vi et lille eksempel til illustration.

EKSEMPEL 3.3. Ved fremstilling af et syntetisk produkt benyttes hovedsagligt to råmaterialer A og B. Kvaliteten af det færdige produkt kan beskrives ved en stokastisk variabel, der følger den normale fordeling med middelværdi μ og varians σ^2. Det vides desuden, at middelværdien afhænger lineært af tilsætningen af A og B, d.v.s.

$$\mu = x_{\mathrm{A}}\theta_{\mathrm{A}} + x_{\mathrm{B}}\theta_{\mathrm{B}},$$

hvor x_{A} angiver den benyttede mængde at materiale A og x_{B} tilsvarende angiver den benyttede mængde af materiale B. σ^2 forudsættes uafhængig af råvaretilsætningen.

Til bestemmelse af θ_{A} og θ_{B} udførtes nu 3 forsøg efter følgende plan

Forsøg	Indhold af A	Indhold af B
1	100%	0%
2	0%	100%
3	50%	50%

De enkelte forsøg antages stokastisk uafhængige. Den simultane fordeling af forsøgsresultaterne Y_1, Y_2, Y_3 er da en tredimensional normal fordeling med middelværdi

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \theta_{\mathrm{A}} \\ \theta_{\mathrm{B}} \end{bmatrix} = \mathbf{x}\theta,$$

og dispersionsmatrix $\sigma^2\mathbf{I}$.

Vi har

$$\mathbf{x}'\mathbf{x} = \begin{bmatrix} \frac{5}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{5}{4} \end{bmatrix} \quad \Rightarrow \quad (\mathbf{x}'\mathbf{x})^{-1} = \begin{bmatrix} \frac{5}{6} & -\frac{1}{6} \\ -\frac{1}{6} & \frac{5}{6} \end{bmatrix},$$

og

$$\mathbf{x}'\mathbf{y} = \begin{bmatrix} y_1 + \frac{1}{2}y_3 \\ y_2 + \frac{1}{2}y_3 \end{bmatrix},$$

hvorfor

$$\begin{bmatrix} \hat{\theta}_A \\ \hat{\theta}_B \end{bmatrix} = \begin{bmatrix} \frac{5}{6} & -\frac{1}{6} \\ -\frac{1}{6} & \frac{5}{6} \end{bmatrix} \begin{bmatrix} y_1 + \frac{1}{2}y_3 \\ y_2 + \frac{1}{2}y_3 \end{bmatrix} = \begin{bmatrix} \frac{5}{6}y_1 - \frac{1}{6}y_2 + \frac{1}{3}y_3 \\ -\frac{1}{6}y_1 + \frac{5}{6}y_2 + \frac{1}{3}y_3 \end{bmatrix}.$$

I det aktuelle tilfælde observeredes

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 10.11 \\ 0.81 \\ 5.24 \end{bmatrix},$$

således at

$$\begin{bmatrix} \hat{\theta}_A \\ \hat{\theta}_B \end{bmatrix} = \begin{bmatrix} 10.037 \\ 0.735 \end{bmatrix}.$$

Heraf findes let

$$\hat{\mathrm{E}}(\mathbf{Y}) = \mathbf{x}\hat{\theta} = \begin{bmatrix} 10.037 \\ 0.735 \\ 5.386 \end{bmatrix},$$

og

$$\mathbf{Y} - \hat{\mathrm{E}}(\mathbf{Y}) = \mathbf{Y} - \mathbf{x}\hat{\theta} = \begin{bmatrix} 0.07 \\ 0.07 \\ -0.15 \end{bmatrix}.$$

Dette giver residualkvadratafvigelsessummen

$$(\mathbf{Y} - \mathbf{x}\hat{\theta})'(\mathbf{Y} - \mathbf{x}\hat{\theta}) = 0.07^2 + 0.07^2 + 0.15^2 = 0.0338,$$

hvorfor et centralt skøn over σ^2 er

$$\frac{1}{3-2}0.0338 = 0.0338.$$

♦

3.1.3 Tilfældet $\mathbf{x}'\Sigma^{-1}\mathbf{x}$ singulær

Hvis $\mathrm{rg}(\mathbf{x}) = p < k$ er $\mathbf{x}'\Sigma^{-1}\mathbf{x}$ singulær, og vi kan derfor ikke umiddelbart løse ligningen.

$$(\mathbf{x}'\Sigma^{-1}\mathbf{x})\hat{\theta} = \mathbf{x}'\Sigma^{-1}\mathbf{y}.$$

Hvis vi kan finde en pseudoinvers til $\mathbf{x}'\Sigma^{-1}\mathbf{x}$ kan vi umiddelbart skrive

$$\hat{\theta} = (\mathbf{x}\Sigma^{-1}\mathbf{x})^{-}\mathbf{x}'\Sigma^{-}\mathbf{y}.$$

Fra tid til anden er det dog muligt at benytte et lille trick ved bestemmelsen af den pseudoinverse. Grunden til singulariteten er, at vi har for mange parametre. Det ville derfor være rimeligt at kræve, at θ kun kan variere frit i et sideunderrum i R^k. Et sådant eg. bestemmes ved, at θ skal tilfredsstille de lineære ligninger (bånd)

$$\mathbf{b}\theta = \mathbf{c}$$

eller

$$\begin{bmatrix} b_{11} & \cdots & b_{1k} \\ \vdots & & \vdots \\ b_{m1} & \cdots & b_{mk} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_k \end{bmatrix} = \begin{bmatrix} c_1 \\ \vdots \\ c_m \end{bmatrix}.$$

Hvis der overhovedet eksisterer θ'er, der tilfredsstiller dette ligningssystem, da udgør de et sideunderrum af dimension $k - \mathrm{rg}(\mathbf{b})$.

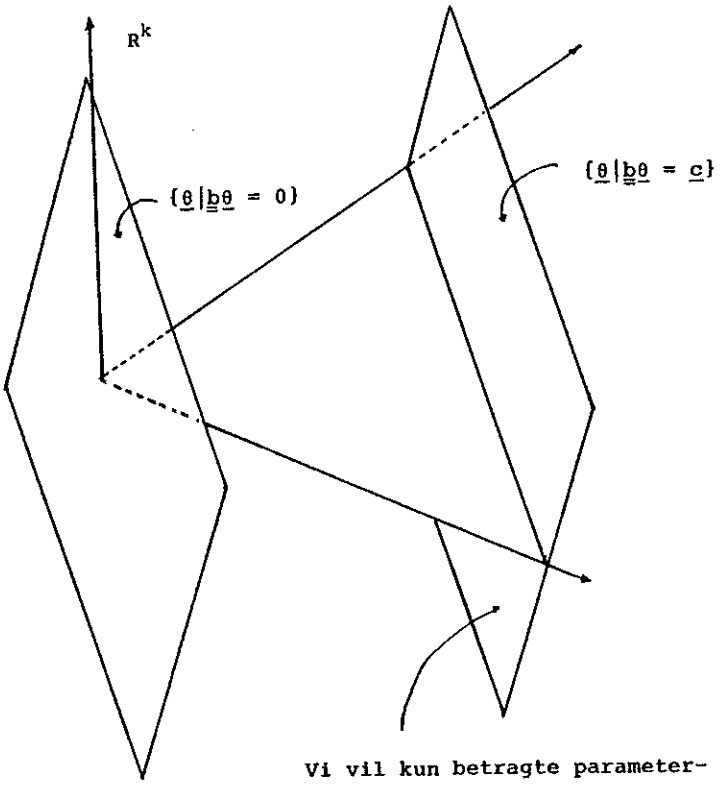
Da $\mathrm{rg}(\mathbf{x}) = p$, og da vi har k θ-komponenter, vil det være rimeligt at "fjerne" $k - p$ af disse, d.v.s. at kræve, at $k - \mathrm{rg}(\mathbf{b}) = p$ eller $k = p + \mathrm{rg}(\mathbf{b})$.

Hvis nu

$$\mathrm{rg}\begin{bmatrix} \mathbf{x} \\ \mathbf{b} \end{bmatrix} = \mathrm{rg}\begin{bmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nk} \\ b_{11} & \cdots & b_{1k} \\ \vdots & & \vdots \\ b_{m1} & \cdots & b_{mk} \end{bmatrix} = k,$$

kan vi betragte "modellen"

$$\begin{bmatrix} \mathbf{Y} \\ \mathbf{c} \end{bmatrix} = \begin{bmatrix} \mathbf{x} \\ \mathbf{b} \end{bmatrix}\theta + \begin{bmatrix} \varepsilon \\ \mathbf{0} \end{bmatrix}.$$



Vi vil kun betragte parameter-værdier $\underline{\theta}$, der ligger i dette sideunderrum i R^k.

Vi sætter

$$D = \begin{bmatrix} \Sigma^{-1} & \mathbf{0}_{n,m} \\ \mathbf{0}_{m,n} & \mathbf{I}_{m,m} \end{bmatrix} = \begin{bmatrix} \Sigma^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix},$$

hvor den forkortede skrivemåde ikke vil give anledning til vanskeligheder.

Beregner vi nu helt sædvanligt

$$\begin{aligned} \hat{\theta} &= \{[\mathbf{x}'\mathbf{b}']\mathbf{D}\begin{bmatrix} \mathbf{x} \\ \mathbf{b} \end{bmatrix}\}^{-1}\{[\mathbf{x}'\mathbf{b}']\mathbf{D}\begin{bmatrix} \mathbf{y} \\ \mathbf{c} \end{bmatrix}\} \\ &= \{\mathbf{x}'\Sigma^{-1}\mathbf{x} + \mathbf{b}'\mathbf{b}\}^{-1}\{\mathbf{x}'\Sigma^{-1}\mathbf{y} + \mathbf{b}'\mathbf{c}\}, \end{aligned}$$

har vi fået en størrelse, der minimaliserer

$$\begin{aligned} \mathrm{g}(\theta) &= \{\begin{bmatrix} \mathbf{y} \\ \mathbf{c} \end{bmatrix} - \begin{bmatrix} \mathbf{x} \\ \mathbf{b} \end{bmatrix}\theta\}'\mathbf{D}\{\begin{bmatrix} \mathbf{y} \\ \mathbf{c} \end{bmatrix} - \begin{bmatrix} \mathbf{x} \\ \mathbf{b} \end{bmatrix}\theta\} \\ &= \begin{bmatrix} \mathbf{y} - \mathbf{x}\theta \\ \mathbf{0} \end{bmatrix}'\begin{bmatrix} \Sigma^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}\begin{bmatrix} \mathbf{y} - \mathbf{x}\theta \\ \mathbf{0} \end{bmatrix} \\ &= (\mathbf{y} - \mathbf{x}\theta)'\Sigma^{-1}(\mathbf{y} - \mathbf{x}\theta) \\ &= \|\mathbf{y} - \mathbf{x}\theta\|^2. \end{aligned}$$

Da dette netop er denne størrelse vi skal bestemme for at finde ML-skønnene, ser vi, at

$$\hat{\theta} = \{\mathbf{x}'\Sigma^{-1}\mathbf{x} + \mathbf{b}'\mathbf{b}\}^{-1}\{\mathbf{x}'\Sigma^{-1}\mathbf{y} + \mathbf{b}'\mathbf{c}\}$$

virkelig er **maximum likelihood estimatoren for θ**. Det eneste der kræves, er altså, at vi kan finde en matrix $\mathbf{b}$ så $\begin{bmatrix} \mathbf{x} \\ \mathbf{b} \end{bmatrix}$ får fuld rang, og dette svarer til, at vi indskrænker θ's variationsområde.

Dispersionsmatricen for $\hat{\theta}$ bliver

$$\mathrm{D}(\hat{\theta}) = \sigma^2\{\mathbf{x}'\Sigma^{-1}\mathbf{x} + \mathbf{b}'\mathbf{b}\}^{-1}\mathbf{x}'\Sigma^{-1}\mathbf{x}\{\mathbf{x}'\Sigma^{-1}\mathbf{x} + \mathbf{b}'\mathbf{b}\}^{-1}.$$

Dette udtryk fås ved umiddelbar anvendelse af sætning 2.5

Det centrale skøn over σ^2 bliver som før

$$\frac{1}{n - \mathrm{rg}\,\mathbf{x}}\|\mathbf{y} - \mathbf{x}\hat{\theta}\|^2$$

Her er $n - \mathrm{rg}\,\mathbf{x} = n - k + \mathrm{rg}\,\mathbf{b}$.

Vi giver nu først et lille teoretisk

EKSEMPEL 3.4. Vi betragter en meget simpel ensidet variansanalyse med to grupper med to observationer i hver. Vi kan forestille os, at man vil undersøge en katalysators effekt på udbyttet af en proces, og at man derfor foretager 4 forsøg, to med katalysatoren i niveau A og to med katalysatoren i niveau B. Vi har derfor målinger

niveau A: Y_{11}, Y_{12}
niveau B: Y_{21}, Y_{22}

Hvis vi regner med, at observationerne er stokastisk uafhængige og har middelværdier

$$\begin{aligned} \mathrm{E}(Y_{11}) &= \mathrm{E}(Y_{12}) = \theta_1 \\ \mathrm{E}(Y_{21}) &= \mathrm{E}(Y_{22}) = \theta_2, \end{aligned}$$

kan vi skrive modellen som

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} + \varepsilon = \mathbf{x}\theta + \varepsilon.$$

Man finder let

$$\mathbf{x}'\mathbf{x} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix},$$

og

$$\hat{\theta} = \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} y_{11} + y_{12} \\ y_{21} + y_{22} \end{bmatrix} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \end{bmatrix},$$

hvilket er de sædvanligt kendte estimatorer. Anvender vi i stedet parametriseringen

$$\begin{aligned} \mathrm{E}(Y_{11}) &= \mathrm{E}(Y_{12}) = \mu + \alpha_1 \\ \mathrm{E}(Y_{21}) &= \mathrm{E}(Y_{22}) = \mu + \alpha_2 \end{aligned}$$

d.v.s. vi udtrykker effekten af en katalysator som et niveau plus den specifikke effekt af den pågældende katalysator - da fås

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{bmatrix} + \varepsilon = \mathbf{x}\alpha + \varepsilon.$$

Det ses let, at $\mathbf{x}$ har rangen 2 (summen af de to sidste søjler er lig den første). Vi vil derfor forsøge at indføre et lineært bånd mellem parametrene. Vi forsøger med

$$\alpha_1 + \alpha_2 = 0 \quad \text{d.v.s.:} \quad \begin{pmatrix} 0 & 1 & 1 \end{pmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{bmatrix} = 0.$$

Vi opstiller nu formelt modellen

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{bmatrix} + \begin{bmatrix} \varepsilon \\ 0 \end{bmatrix},$$

eller

$$\begin{bmatrix} \mathbf{Y} \\ 0 \end{bmatrix} = \begin{bmatrix} \mathbf{x} & \\ 0 & 1 \quad 1 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha_1 \\ \alpha_2 \end{bmatrix} + \begin{bmatrix} \varepsilon \\ 0 \end{bmatrix}.$$

Vi har herefter

$$\begin{bmatrix} \mathbf{x} & \\ 0 & 1 \quad 1 \end{bmatrix}' \begin{bmatrix} \mathbf{x} & \\ 0 & 1 \quad 1 \end{bmatrix} = \mathbf{x}'\mathbf{x} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 3 & 1 \\ 2 & 1 & 3 \end{bmatrix}.$$

Den inverse til denne matrix er

$$\begin{bmatrix} \frac{1}{2} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{2} & 0 \\ -\frac{1}{4} & 0 & \frac{1}{2} \end{bmatrix}.$$

Da

$$\begin{bmatrix} \mathbf{x} & \\ 0 & 1 \quad 1 \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ 0 \end{bmatrix} = \begin{bmatrix} \sum y_{ij} \\ y_{11} + y_{12} \\ y_{21} + y_{22} \end{bmatrix},$$

fås

$$\begin{bmatrix} \hat{\mu} \\ \hat{\alpha_1} \\ \hat{\alpha_2} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{2} & 0 \\ -\frac{1}{4} & 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \sum y_{ij} \\ y_{11} + y_{12} \\ y_{21} + y_{22} \end{bmatrix} = \begin{bmatrix} \bar{y} \\ \bar{y}_1 - \bar{y} \\ \bar{y}_2 - \bar{y} \end{bmatrix},$$

d.v.s. netop de estimatorer vi er vant til. (NB. Vi ved på forhånd, at vi vil komme frem til disse, jvf. p. 122). ♦

Vi giver nu et lidt mere praktisk betonet eksempel på estimation af parametre i tilfælde, hvor $\mathbf{x}'\Sigma^{-1}\mathbf{x}$ er singulær.

EKSEMPEL 3.5. Til fremstilling af enzymer kan anvendes 2 principielt forskellige bakterietyper. Den ene udskiller gennem sit stofskifte syre under dyrkning (syredanner), den anden udskiller neutrale stofskifteprodukter. For at regulere pH-værdien i det substrat, bakterierne dyrkes på, kan der tilsættes en såkaldt pH-buffer. Man ved, at pH-bufferen i sig selv ikke har nogen effekt på enzymudbyttet, men udelukkende virker gennem en vekselvirkning med surhedsgraden og bakteriernes stofskifteprodukter.

For en "neutral" bakteriestamme, som lever på et substrat uden pH-buffer, kendes middelenzymudbytteet (normudbyttet). For at estimere de ovennævnte vekselvirkningseffekter er ved 7 forsøg målt forskellen mellem normudbyttet og det faktiske enzymudbytte som nedenfor angivet.

		pH-buffer	
		tilsat	ikke tilsat
bakterie	syredanner	0,-2	-19,-15
stamme	neutral	-6, 0,-2	

Tabel 3.1: Forskelle mellem normudbytte og det faktiske enzymudbytte ved forskellige forsøgsomstændigheder.

Vi vil først formulere en matematisk model, der kan beskrive ovenstående eksperiement.

Der foreligger observationer

$$\begin{aligned} &y_{11\nu}, && \nu = 1,2 \\ &y_{12\nu}, && \nu = 1,2 \\ &y_{21\nu}, && \nu = 1,2,3. \end{aligned}$$

Disse antages at have forventningsværdierne

$$\begin{aligned} \mathrm{E}(y_{11\nu}) &= \mu_1 + \theta_{11} \\ \mathrm{E}(y_{12\nu}) &= \mu_1 + \theta_{12} \\ \mathrm{E}(y_{21\nu}) &= \theta_{21}, \end{aligned}$$

hvor μ_1 er effekten af at anvende syredannende bakterier og θ_{ij} angiver vekselvirkningerne mellem tilstedeværelsen af pH-buffer og bakteriestammetypen.

Endvidere antages observationerne at være stokastiske uafhængige og at have samme ukendte varians σ^2.

Vi kan nu formulere modellen som en generel lineær model. Vi har

$$\begin{bmatrix} Y_{111} \\ Y_{112} \\ Y_{121} \\ Y_{122} \\ Y_{211} \\ Y_{212} \\ Y_{213} \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \theta_{11} \\ \theta_{12} \\ \theta_{21} \end{bmatrix} + \varepsilon,$$

hvor fejlen $\varepsilon \in \mathrm{N}_7(\mathbf{0}, \sigma^2\mathbf{I})$.

Vi finder

$$\mathbf{x}'\mathbf{x} = \begin{bmatrix} 4 & 2 & 2 & 0 \\ 2 & 2 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix},$$

og

$$\mathbf{x}'\mathbf{y} = \begin{bmatrix} y_{1..} \\ y_{11.} \\ y_{12.} \\ y_{21.} \end{bmatrix},$$

hvor et punktum på en indeksplads angiver, at der er summeret over det pågældende indeks.

Da $\mathbf{x}'\mathbf{x}$ åbenbart kun har rangen 3, kan vi ikke umiddelbart invertere den. I stedet kan vi finde en pseudoinvers. Vi benytter sætning 1.7 p. 26 og får

$$(\mathbf{x}'\mathbf{x})^- = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{3} \end{bmatrix},$$

hvorfor estimatorerne for parametrene bliver - med dette specielle valg af pseudoinvers -

$$\hat{\theta} = (\mathbf{x}'\mathbf{x})^-\mathbf{x}'\mathbf{y} = \begin{bmatrix} 0 \\ \bar{y}_{11.} \\ \bar{y}_{12.} \\ \bar{y}_{21.} \end{bmatrix},$$

hvor f.eks.

$$\bar{y}_{21.} = \frac{1}{3}\sum_{\nu=1}^{3} y_{21\nu}.$$

Da

$$\mathbf{I} - (\mathbf{x}'\mathbf{x})^{-}\mathbf{x}'\mathbf{x} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

er

$$(\mathbf{I} - (\mathbf{x}'\mathbf{x})^{-}\mathbf{x}'\mathbf{x})\mathbf{z} = \begin{bmatrix} z_1 \\ -z_1 \\ -z_1 \\ 0 \end{bmatrix}$$

Ifølge sætning 1.6 er den fuldstændige løsning til normalligningerne derfor alle vektorer af formen

$$\hat{\theta} + \begin{bmatrix} t \\ -t \\ -t \\ 0 \end{bmatrix} = \begin{bmatrix} t \\ \bar{y}_{11.} - t \\ \bar{y}_{12.} - t \\ \bar{y}_{21.} \end{bmatrix}, \qquad t \in R.$$

En vilkårlig maksimum likelihood estimator for θ er altså af denne form.

Den observerede værdi af $\hat{\theta}$ er

$$\hat{\theta}_{\text{obs}} = \begin{bmatrix} 0 \\ -1 \\ -17 \\ -2\frac{2}{3} \end{bmatrix}.$$

Det er åbenbart, at denne estimator ikke er særligt tilfredsstillende, idet $\hat{\mu_1}$ f.eks. altid vil være 0. For at få estimatorer, der bedre svarer til vore forestillinger om den fysiske virkelighed, må vi lægge nogle bånd på parametrene. Det forekommer rimeligt at kræve

$$\theta_{11} + \theta_{12} = 0,$$

d.v.s.

$$\begin{pmatrix} 0 & 1 & 1 & 0 \end{pmatrix} \begin{bmatrix} \mu_1 \\ \theta_{11} \\ \theta_{12} \\ \theta_{21} \end{bmatrix} = 0,$$

eller

$$\mathbf{b}\,\theta = 0.$$

Det er åbenbart, at

$$\mathrm{rg}(\begin{bmatrix} \mathbf{x} \\ \mathbf{b} \end{bmatrix}) = 4,$$

hvorfor vi kan anvende resultatet fra p. 122. Vi finder

$$\begin{aligned} \mathbf{x}'\mathbf{x} + \mathbf{b}'\mathbf{b} &= \begin{bmatrix} 4 & 2 & 2 & 0 \\ 2 & 2 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \\ &= \begin{bmatrix} 4 & 2 & 2 & 0 \\ 2 & 3 & 1 & 0 \\ 2 & 1 & 3 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}. \end{aligned}$$

Da

$$\begin{bmatrix} 4 & 2 & 2 \\ 2 & 3 & 1 \\ 2 & 1 & 3 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{2} & 0 \\ -\frac{1}{4} & 0 & \frac{1}{2} \end{bmatrix},$$

fås

$$(\mathbf{x}'\mathbf{x} + \mathbf{b}'\mathbf{b})^{-1} = \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} & -\frac{1}{4} & 0 \\ -\frac{1}{4} & \frac{1}{2} & 0 & 0 \\ -\frac{1}{4} & 0 & \frac{1}{2} & 0 \\ 0 & 0 & 0 & \frac{1}{3} \end{bmatrix}.$$

Vi finder nu

$$\hat{\theta} = (\mathbf{x}'\mathbf{x} + \mathbf{b}'\mathbf{b})^{-1}\mathbf{x}'\mathbf{y} = \begin{bmatrix} \bar{y}_{1\cdot} \\ \bar{y}_{11\cdot} - \bar{y}_{1\cdot\cdot} \\ \bar{y}_{12\cdot} - \bar{y}_{1\cdot\cdot} \\ \bar{y}_{21\cdot} \end{bmatrix}.$$

Den observerede værdi er

$$\begin{bmatrix} -9 \\ 8 \\ -8 \\ -2\frac{2}{3} \end{bmatrix} \quad \left(= \begin{bmatrix} \text{syredanneeffekt} \\ \text{vex.v. buffer \& syre} \\ \text{vex.v. (-buffer) \& syre} \\ \text{vex.v. buffer \& neutral} \end{bmatrix} \right).$$

Vi finder dernæst dispersionsmatricen for $\hat{\theta}$. Der gælder

$$\begin{aligned} \mathrm{D}(\hat{\theta}) &= \sigma^2(\mathbf{x}'\mathbf{x}+\mathbf{b}'\mathbf{b})^{-1}\mathbf{x}'\mathbf{x}(\mathbf{x}'\mathbf{x}+\mathbf{b}'\mathbf{b})^{-1} \\ &= \sigma^2 \begin{bmatrix} \frac{1}{4} & 0 & 0 & 0 \\ 0 & \frac{1}{4} & -\frac{1}{4} & 0 \\ 0 & -\frac{1}{4} & \frac{1}{4} & 0 \\ 0 & 0 & 0 & \frac{1}{3} \end{bmatrix}, \end{aligned}$$

d.v.s. estimatorerne er ikke uafhængige.

For at estimere σ^2 finder vi residualvektoren. Da

$$\mathbf{x}\hat{\theta} = \begin{bmatrix} \hat{\mu}_1+\hat{\theta}_{11} \\ \hat{\mu}_1+\hat{\theta}_{11} \\ \hat{\mu}_1+\hat{\theta}_{12} \\ \hat{\mu}_1+\hat{\theta}_{12} \\ \hat{\theta}_{21} \\ \hat{\theta}_{21} \\ \hat{\theta}_{21} \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \\ -17 \\ -17 \\ -2\frac{2}{3} \\ -2\frac{2}{3} \\ -2\frac{2}{3} \end{bmatrix},$$

er **residualvektoren**

$$\mathbf{y}-\mathbf{x}\hat{\theta} = \begin{bmatrix} 1 \\ -1 \\ -2 \\ 2 \\ -3\frac{1}{3} \\ 2\frac{2}{3} \\ \frac{2}{3} \end{bmatrix}.$$

Dermed er

$$||\mathbf{y}-\mathbf{x}\hat{\theta}||^2 = (\mathbf{y}-\mathbf{x}\hat{\theta})'(\mathbf{y}-\mathbf{x}\hat{\theta}) = 1^2+\cdots+(\frac{2}{3})^2 = 28\frac{2}{3}.$$

Et centralt skøn over σ^2 er derfor

$$s^2 = \frac{1}{7-3}28\frac{2}{3} = 7\frac{1}{6}.$$

♦

3.1.4 Estimation under bibetingelser

Vi skal nu beskæftige os med et problem, der meget minder om det i foregående afsnit behandlede, nemlig estimation af parametre, der skal tilfredsstille lineære bånd som

$$\mathbf{H}'\theta = \xi.$$

Dette vil f.eks. være tilfældet, hvor man estimerer vinkler i en bestemt trekant. De skal tilfredsstille

$$\begin{pmatrix} 1 & 1 & 1 \end{pmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} = 180°,$$

og det vil derfor være rimeligt også at kræve dette opfyldt af estimaterne.

Hovedresultatet om estimation af θ kan udtrykkes i

SÆTNING 3.3. Lad $\mathrm{E}(\mathbf{Y}) = \mathbf{x}\theta$, hvor $\mathbf{Y}$ er en n-dimensional vektor, $\mathbf{x}$ en $n \times k$ matrix af kendte koefficienter, og θ den k-dimensionale vektor af ukendte parametre der yderligere forudsættes at tilfredsstille de s lineære restriktioner

$$\mathbf{H}'\theta = \xi,$$

hvor $\mathbf{H}$ er en $k \times s$ matrix, og ξ en s-dimensional vektor. Endelig forudsættes, at $\mathrm{D}(\mathbf{Y}) = \sigma^2\Sigma$, hvor Σ er en kendt matrix. Mindste kvadraters estimator $\tilde{\theta}$ for θ under bibetingelsen $H'\theta = \xi$ er da løsning til ligningssystemet

$$\begin{bmatrix} \mathbf{x}'\Sigma^{-1}\mathbf{x} & \mathbf{H} \\ \mathbf{H}' & \mathbf{0} \end{bmatrix} \begin{bmatrix} \theta \\ \lambda \end{bmatrix} = \begin{bmatrix} \mathbf{x}'\Sigma^{-1}\mathbf{y} \\ \xi \end{bmatrix}.$$

▲

BEVIS 3.3. Vi skal finde

$$\min_{\mathbf{H}'\theta=\xi} (\mathbf{Y} - \mathbf{x}\theta)'\Sigma^{-1}(\mathbf{Y} - \mathbf{x}\theta).$$

Vi indfører Lagrangemultiplikatorer λ og sætter

$$\mathrm{F}(\theta, \lambda) = \frac{1}{2}(\mathbf{Y} - \mathbf{x}\theta)'\Sigma^{-1}(\mathbf{Y} - \mathbf{x}\theta) + \lambda'(\mathbf{H}'\theta - \xi).$$

Da bliver

$$\begin{aligned}\frac{\partial F}{\partial \theta} &= -\mathbf{x}'\Sigma^{-1}\mathbf{y} + \mathbf{x}'\Sigma^{-1}\mathbf{x}\theta + \mathbf{H}\lambda \\ \frac{\partial F}{\partial \lambda} &= \mathbf{H}'\theta - \xi.\end{aligned}$$

Disse to afledede er 0 i ethvert ekstremum for $(\mathbf{Y} - \mathbf{x}\theta)'\Sigma^{-1}(\mathbf{Y} - \mathbf{x}\theta)$ under betingelsen $\mathbf{H}'\theta = \xi$. Udnyttes dette fås umiddelbart påstanden i sætningen. ■

Vi betragter dernæst problemet med estimation af σ^2 i

SÆTNING 3.4. Lad en pseudoinvers til koefficientmatricen i sætning 3.3 være

$$\begin{bmatrix} \mathbf{x}'\Sigma^{-1}\mathbf{x} & \mathbf{H} \\ \mathbf{H}' & \mathbf{0} \end{bmatrix}^- = \begin{bmatrix} \mathbf{C}_1 & \mathbf{C}_2 \\ \mathbf{C}_3 & \mathbf{C}_4 \end{bmatrix}.$$

Da er

$$\mathrm{D}(\tilde{\theta}) = \sigma^2\mathbf{C}_1,$$

og et centralt skøn for σ^2 er

$$\hat{\sigma}^2 = \frac{1}{\mathrm{f}}(\mathbf{Y}'\Sigma^{-1}\mathbf{Y} - \tilde{\theta}'\mathbf{x}'\Sigma^{-1}\mathbf{Y} - \xi'\tilde{\lambda}),$$

hvor $(\tilde{\theta}', \tilde{\lambda}')'$ er løsning til ligningssystemet i sætning 3.3, og

$$\mathrm{f} = n - \mathrm{rg}\begin{pmatrix} \mathbf{x}' \\ \mathbf{H} \end{pmatrix} + \mathrm{rg}(\mathbf{H}).$$

▲

BEVIS 3.4. Ved at indføre den pseudoinverse fås, at

$$\begin{aligned}\tilde{\theta} &= \mathbf{C}_1\mathbf{x}'\Sigma^{-1}\mathbf{Y} + \mathbf{C}_2\xi \\ \tilde{\lambda} &= \mathbf{C}_3\mathbf{x}'\Sigma^{-1}\mathbf{Y} + \mathbf{C}_4\xi.\end{aligned}$$

Heraf fås umiddelbart

$$\begin{aligned}\mathrm{D}(\tilde{\theta}) &= \sigma^2\mathbf{C}_1\mathbf{x}'\Sigma^{-1}\Sigma\Sigma^{-1}\mathbf{x}\mathbf{C}_1' \\ &= \sigma^2\mathbf{C}_1\mathbf{x}'\Sigma^{-1}\mathbf{x}\mathbf{C}_1' \\ &= \sigma^2\mathbf{C}_1\end{aligned}$$

Det sidste lighedstegn følger ved at benytte egenskaber ved pseudoinverse matricer.

Ved direkte indsætning ses, at

$$(\mathbf{Y} - \mathbf{x}\tilde{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{x}\tilde{\theta}) = (\mathbf{Y}'\boldsymbol{\Sigma}^{-1}\mathbf{Y} - \tilde{\theta}'\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{Y} - \xi'\tilde{\lambda}),$$

så vi mangler blot at vise, at antallet af frihedsgrader er det i sætningen anførte. Betragter vi løsningerne til ligningen

$$\mathbf{H}'\theta = \xi,$$

kan disse skrives på formen

$$\theta = \theta_0 + \mathbf{B}\,\beta,$$

hvor θ_0 er en partikulær løsning, og $\mathbf{B}$ er en $(k \times s)$ matrix $(\mathrm{rg}(\mathbf{H}) = k - s)$ med

$$\mathbf{H}'\mathbf{B} = \mathbf{0}.$$

Endelig er β en s-dimensional vektor af "frie", nye parametre. Betragtes

$$\mathbf{Z} = \mathbf{Y} - \mathbf{x}\theta_0,$$

fås

$$\begin{aligned} \mathrm{E}(\mathbf{Z}) &= \mathbf{x}\theta - \mathbf{x}\theta_0 = \mathbf{x}(\theta - \theta_0) \\ &= \mathbf{x}(\theta - \theta_0) \\ &= \mathbf{x}\mathbf{B}\,\beta. \end{aligned}$$

Vi kan nu blot betragte modellen

$$\mathbf{Z} = \mathbf{x}\mathbf{B}\,\beta + \varepsilon,$$

hvor ε er fejlvektoren og løse denne. Dette giver de i det tidligere anførte estimater.

Med

$$\hat{\beta} = (\mathbf{B}'\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x}\,\mathbf{B})^{-}\mathbf{B}'\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{Y},$$

bliver

$$\tilde{\theta} = \theta_0 + \mathbf{B}\,\hat{\beta},$$

hvorfor

$$\begin{aligned}\mathbf{Y} - \mathrm{x}\tilde{\theta} &= \mathbf{Z} + \mathrm{x}\,\theta_0 - \mathrm{x}\,\theta_0 - \mathrm{x}\,\mathbf{B}\,\hat{\beta} \\ &= \mathbf{Z} - \mathrm{x}\,\mathbf{B}\hat{\beta},\end{aligned}$$

og fra den almindelige teori fås at antallet af frihedsgrader bliver $n - \mathrm{rg}(\mathbf{x}\,\mathbf{B})$. Nu er

$$\begin{aligned}\mathrm{rg}(\mathbf{x}\,\mathbf{B}) &= \dim\{\mathbf{x}\,\mathbf{B}\,\beta | \beta \in R^s\} \\ &= \dim\{\mathbf{x}\,\gamma | \mathbf{H}'\gamma = 0, \quad \gamma \in R^m\} \\ &= \mathrm{rg}\begin{pmatrix}\mathbf{x} \\ \mathbf{H}'\end{pmatrix} - \mathrm{rg}(\mathbf{H}).\end{aligned}$$

Det sidste lighedstegn fås af relationen

$$\dim S_1^* + \dim S_2^* = \mathrm{rg}\begin{pmatrix}\mathbf{x} \\ \mathbf{H}'\end{pmatrix},$$

hvor

$$\begin{aligned}S_1^* &= \left\{\begin{pmatrix}\mathbf{x} \\ \mathbf{H}'\end{pmatrix}\gamma \,|\, \gamma \in N(H)\right\} \\ S_1^* &= \left\{\begin{pmatrix}\mathbf{x} \\ \mathbf{H}'\end{pmatrix}\gamma \,|\, \gamma \in N(H)^{\perp}\right\}\end{aligned}$$

(idet man bemærker, at $\dim S_2^* = \mathrm{rg}\,H$). ■

Vi vil nu give et simpelt, illustrativt eksempel.

EKSEMPEL 3.6. Lad os antage, at vi har 3 uafhængige målinger af vinklerne i en trekant (f.eks. ude i terræn), og at disse er fundet til

$$\begin{aligned}v_1 &= 52°, 54° \\ v_2 &= 74°, 74° \\ v_3 &= 48°, 46°.\end{aligned}$$

Vi antager endvidere, at usikkerheden på disse 3 bestemmelser er ens og kan udtrykkes ved en varians σ^2.

Vi formulerer ovenstående som en sædvanlig lineær model med bibetingelse, i.e.

$$\begin{bmatrix} v_{11} \\ v_{12} \\ v_{21} \\ v_{22} \\ v_{31} \\ v_{32} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \epsilon_5 \\ \epsilon_6 \end{bmatrix}$$

$$(1,1,1) \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix} = 180,$$

$$D(\varepsilon) = \sigma^2 \mathbf{I}.$$

Vi finder

$$\begin{bmatrix} \mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x} & \mathbf{H} \\ \mathbf{H}' & \mathbf{0} \end{bmatrix} = \left[\begin{array}{ccc|c} 2 & 0 & 0 & 1 \\ 0 & 2 & 0 & 1 \\ 0 & 0 & 2 & 1 \\ \hline 1 & 1 & 1 & 0 \end{array}\right].$$

En (pseudo)invers til denne er

$$\begin{bmatrix} \mathbf{C}_1 & \mathbf{C}_2 \\ \mathbf{C}_3 & \mathbf{C}_4 \end{bmatrix} = \frac{1}{6} \left[\begin{array}{ccc|c} 2 & -1 & -1 & 2 \\ -1 & 2 & -1 & 2 \\ -1 & -1 & 2 & 2 \\ \hline 2 & 2 & 2 & -4 \end{array}\right].$$

Vi finder derfor

$$\begin{aligned} \tilde{\theta} &= \frac{1}{6} \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix} \begin{bmatrix} 106 \\ 148 \\ 94 \end{bmatrix} + \frac{1}{6} \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix} 180 \\ &= \begin{bmatrix} -5 \\ 16 \\ -11 \end{bmatrix} + \begin{bmatrix} 60 \\ 60 \\ 60 \end{bmatrix} \\ &= \begin{bmatrix} 55 \\ 76 \\ 49 \end{bmatrix}. \end{aligned}$$

Vi bemærker trivielt, at summen af koordinaterne er 180.

Dispersionsmatricen bliver

$$\mathrm{D}(\tilde{\theta}) = \sigma^2 \mathbf{C}_1 = \frac{\sigma^2}{6} \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix}.$$

Skønnet over σ^2 bliver

$$\sigma^2 = \frac{1}{6-3+1}(20992 - 21684 - (-720)) = 7 = 2.6^2,$$

idet

$$\tilde{\lambda} = \frac{1}{6}[2,2,2] \begin{bmatrix} 106 \\ 148 \\ 94 \end{bmatrix} + (-\frac{4}{6})180 = 116 - 120 = -4$$

♦

BEMÆRKNING 3.3. Som det er antydet i ovenstående eksempel nyder denne teori speciel anvendelse inden for **geodæsien** og **landmålingen**. Vi skal dog ikke komme ind på nærmere detaljer her. ▼

3.1.5 Konfidensintervaller for forudsagte værdier. Prediktionsinterval

Vi betragter modellen $(n > k)$

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

hvor

$$\varepsilon \in \mathrm{N}(\mathbf{0}, \sigma^2 \boldsymbol{\Sigma}).$$

Vi vil her benævne Y'erne de afhængige variable og x'erne de uafhængige variable.

Her er som vanligt σ^2 ukendt og $\boldsymbol{\Sigma}$ kendt. Vi har estimatoren

$$\hat{\theta} = (\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x})^{-1}\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}$$

for θ, og σ^2 estimeres ved

$$\begin{aligned}\hat{\sigma^2} &= s^2 = \frac{1}{n-k}\|\mathbf{Y} - \mathbf{x}\hat{\theta}\|^2 \\ &= \frac{1}{n-k}(\mathbf{Y} - \mathbf{x}\hat{\theta})'\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{x}\hat{\theta}).\end{aligned}$$

Hvis vi ønsker at forudsige forventningsværdien af en kommende observation Y af den afhængige variabel, svarende til værdierne

$$(z_1, \ldots, z_k) = \mathbf{z}'$$

af de uafhængige variable, er det åbenbart, at vi vil anvende

$$Z = (z_1, \ldots, z_k)\begin{bmatrix}\hat{\theta}_1 \\ \vdots \\ \hat{\theta}_k\end{bmatrix} = \mathbf{z}'\hat{\theta}$$

som et "bedste" gæt.

Vi har, at $\mathrm{E}(Z) = \mathrm{E}(Y)$, og

$$\begin{aligned}V(Z) &= \mathbf{z}'\,\mathrm{D}(\hat{\theta})\mathbf{z} \\ &= \sigma^2\mathbf{z}'(\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x})^{-1}\mathbf{z} \\ &= \sigma^2 c,\end{aligned}$$

hvor altså

$$c = (z_1, \ldots, z_k)(\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x})^{-1}\begin{bmatrix}z_1 \\ \vdots \\ z_k\end{bmatrix}.$$

Vi får derfor umiddelbart

$$\frac{Z - \mathrm{E}(Y)}{\sigma\sqrt{c}} \in \mathrm{N}(0,1),$$

og dermed

$$\frac{Z - \mathrm{E}(Y)}{S\sqrt{c}} \in \mathrm{t}(n-k).$$

Vi er nu i stand til at formulere og bevise

SÆTNING 3.5. Lad situationen være som ovenfor. Da er $(1-\alpha)$-konfidensintervallet for en kommende observation Y's forventningsværdi

$$[z - \mathrm{t}(n-k)_{1-\frac{\alpha}{2}} s\sqrt{c},\quad z + \mathrm{t}(n-k)_{1-\frac{\alpha}{2}} s\sqrt{c}].$$

▲

BEVIS 3.5. Af ovenstående betragtninger følger umiddelbart

$$1-\alpha = P\{Z - \mathrm{t}(n-k)_{1-\frac{\alpha}{2}} s\sqrt{c} \leq \mathrm{E}(Y) \leq Z + \mathrm{t}(n-k)_{1-\frac{\alpha}{2}} s\sqrt{c}\},$$

og vi har derfor sætningen. ■

Nu er man ofte mere interesseret i et konfidensinterval for kommende observationer end for observationernes forventningsværdi. Vi vælger at betragte det lidt mere generelle problem, at bestemme et konfidensinterval for gennemsnittet $\bar{Y}_q$ af q observationer taget under betingelsen $(z_1, \ldots, z_k)$. Vi har, hvis $Y_{iq} \in \mathrm{N}(\mathrm{E}(Y), c_1\sigma^2)$, at

$$\bar{Y}_q \in \mathrm{N}(\mathrm{E}(Y), \frac{c_1}{q}\sigma^2).$$

Hvis vi nu forudsætter, at de kommende observationer er uafhængige af dem, vi har fået, er

$$Z - \bar{Y}_q \in \mathrm{N}(0, \sigma^2(c + \frac{c_1}{q})),$$

d.v.s.

$$\frac{Z - \bar{Y}_q}{S\sqrt{c + \frac{c_1}{q}}} \in \mathrm{t}(n-k).$$

Heraf udledes som før

SÆTNING 3.6. Lad os forudsætte at q kommende observationer underbetingelserne $(z_1, \ldots, z_k)$ har varians $c_1\sigma^2$, og at de er indbyrdes uafhængige og uafhængige af de tidligere observationer. Da er $(1-\alpha)$-konfidensintervallet for gennemsnittet af de q målinger lig intervallet

$$[z - \mathrm{t}(n-k)_{1-\frac{\alpha}{2}} s\sqrt{c + \frac{c_1}{q}}, z + \mathrm{t}(n-k)_{1-\frac{\alpha}{2}} s\sqrt{c + \frac{c_1}{q}}].$$

▲

BEMÆRKNING 3.4. Ovenstående interval er et konfidensinterval for en observation, og ikke, som vi er vant til, for en parameter. Man taler derfor hyppigt om et **prediktionsinterval** for at skelne de to situationer fra hinanden. ▼

BEMÆRKNING 3.5. Vi ser, at overgangen til intervallet for $\bar{Y}_q$ i stedet for intervallet for $\mathrm{E}(\bar{Y}_q) = \mathrm{E}(Y)$ blot består i, at leddet under rodtegnet er øget med variansen på $\frac{\bar{Y}_q}{\sigma}$. ▼

EKSEMPEL 3.7. Vi betragter følgende samhørende målinger af en uafhængig variabel x og en afhængig variabel y:

x	0	1	2	3	4	5	6
y	0.4	0.3	1.5	1.3	1.9	4.2	8

Vi antager, at y'erne er realiserede udfald af uafhængige stokastiske variable $Y_1, \ldots, Y_7$, der er normalt fordelte med middelværdier

$$\mathrm{E}(Y|x) = \beta x^2$$

og varianser

$$\mathrm{V}(Y|0) = \sigma^2, \quad \mathrm{V}(Y|x) = x^2\sigma^2, \qquad x > 0.$$

Vi søger en konfidensinterval for en kommende måling svarende til $x = 10$. Denne måling kaldes Y, og vi har

$$\begin{aligned} \mathrm{E}(Y) &= 100\beta \\ \mathrm{V}(Y) &= 100\sigma^2 \end{aligned}$$

Vi omformulerer nu problemet til matrix form:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 4 \\ 9 \\ 16 \\ 25 \\ 36 \end{bmatrix} \beta + \begin{bmatrix} \epsilon_1 \\ \cdot \\ \cdot \\ \cdot \\ \\ \\ \epsilon_7 \end{bmatrix} = \mathbf{x}\beta + \epsilon,$$

hvor

$$\mathrm{D}(\epsilon) = \sigma^2 \begin{bmatrix} 1 & & \cdot & \cdot & \cdot & & 0 \\ & 1 & & & & & \\ \cdot & & 4 & & & & \cdot \\ \cdot & & & 9 & & & \cdot \\ \cdot & & & & 16 & & \cdot \\ & & & & & 25 & \\ 0 & & \cdot & \cdot & \cdot & & 36 \end{bmatrix} = \sigma^2 \boldsymbol{\Sigma}.$$

Vi har, at

$$\begin{aligned} \mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x} &= (0,1,4,9,16,25,36)\,\mathrm{diag}(1,1,\frac{1}{4},\ldots,\frac{1}{36}) \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 36 \end{bmatrix} \\ &= 91. \\ \mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{y} &= 0.3 + 1.5 + 1.3 + 1.9 + 4.2 + 8.0 = 17.2. \end{aligned}$$

Altså er

$$\hat{\beta} = \frac{17.2}{91} = 0.1890,$$

og

$$P_M(\mathbf{y}) = \begin{bmatrix} 0 \\ 1 \\ 4 \\ 9 \\ 16 \\ 25 \\ 36 \end{bmatrix} \cdot 0.1890 = \begin{bmatrix} 0 \\ 0.1890 \\ 0.7560 \\ 1.7010 \\ 3.0240 \\ 4.7250 \\ 6.8040 \end{bmatrix}.$$

Residualerne er

$$\mathbf{y} - P_M(\mathbf{y}) = \begin{bmatrix} 0.4000 \\ 0.7440 \\ 0.1110 \\ -0.4010 \\ -1.1240 \\ -0.5250 \\ 1.1960 \end{bmatrix},$$

hvorfor

$$\begin{aligned}\|\mathbf{y} - P_M(\mathbf{y})\|^2 &= (0.4000 \cdots 1.1960) \begin{bmatrix} \frac{1}{1} & & & \\ & \frac{1}{1} & & \\ & & \ddots & \\ & & & \frac{1}{36} \end{bmatrix} \begin{bmatrix} 0.4000 \\ \vdots \\ 1.1960 \end{bmatrix} \\ &= 0.45829\end{aligned}$$

d.v.s.

$$\hat{\sigma^2} = s^2 = \frac{1}{7-1} 0.45829 = 0.07638 = 0.27637^2.$$

Konstanterne c og c_1 er lig med

$$\begin{aligned} c &= 100 \cdot \frac{1}{91} \cdot 100 = 109.89 \\ c_1 &= 10^2 = 100. \end{aligned}$$

Forudsigelsen til x lig 10 er

$$z = 100\hat{\beta} = 18.90$$

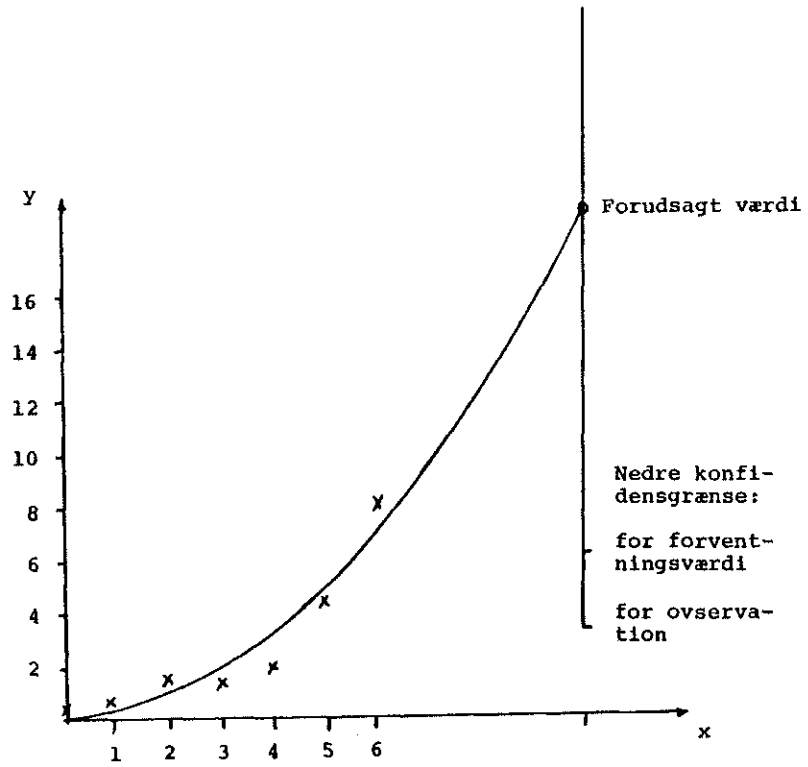
Konfidensintervallet for forventningsværdien i $x = 10$ er derfor givet ved

$$\begin{aligned} & 18.90 \pm \mathrm{t}(6)_{0.975} 0.2764 \sqrt{109.89} \\ = \; & 18.90 \pm 2.447 \cdot 0.2764 \sqrt{109.89} \\ = \; & 18.90 \pm 7.09. \end{aligned}$$

Intervallet for den næste observation er

$$\begin{aligned} & 18.90 \pm \mathrm{t}(6)_{0.975} \cdot 0.2764 \sqrt{109.89 + 100}. \\ = \; & 18.90 \pm 9.80, \end{aligned}$$

d.v.s. et væsentligt bredere interval end for den forventede værdi. Forklaringen er simpelthen, at vi har en varians på $10^2\sigma^2 = 100\sigma^2$ i $x = 10$. Vi afbilder observationerne og det estimerede polynomium i nedenstående graf. Endvidere er de to konfidensintervaller angivet. ♦



3.2 Test i den generelle lineære model

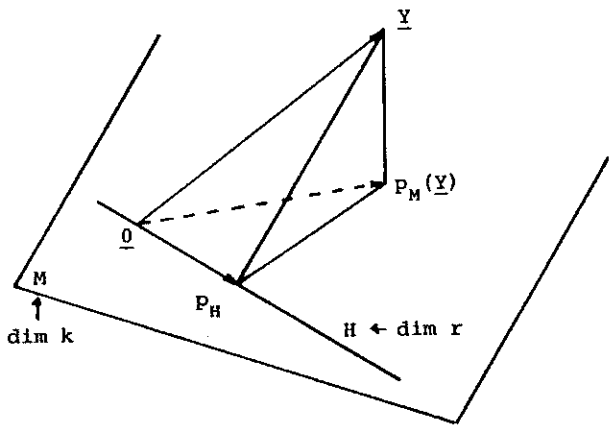
I dette afsnit skal vi dels undersøge om middelværdivektoren kan antages at ligge i et ægte underrum af "modelrummet", og dels undersøge om middelværdivektoren successivt kan antages at ligge i underrum af aftagende dimensioner. Først

3.2.1 Test for lavere dimension af modelrum

Lad $\mathbf{Y} \in \mathrm{N}_n(\mu, \sigma^2\Sigma)$, hvor Σ er regulær og kendt. Vi forudsætter, at $\mu \in M$, et k-dimensionalt underrum, og vi vil teste hypotesen

$$H_0 : \mu \in H \quad \text{mod} \quad H_1 : \mu \in M \backslash H,$$

hvor H er et r-dimensionalt underrum af M. Vi betragter i det følgende den ved Σ^{-1} definerede norm. Maksimum likelihood estimatoren for μ er da $p_M(\mathbf{Y})$-projektionen på M - og hvis H_0 er sand, da er maksimum likelihood estimatoren $p_M(\mathbf{Y})$, $\mathbf{Y}$'s projektion på H. ML estimatorerne for σ^2 er i de to tilfælde $\frac{1}{n}\|\mathrm{y} - p_M(\mathrm{y})\|^2$ henholdsvis $\frac{1}{n}\|\mathrm{y} - p_H(\mathrm{y})\|^2$.



Likelihood funktionerne er

$$\begin{aligned} \mathrm{L}(\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi}^n}\frac{1}{\sigma^n}\frac{1}{\sqrt{\det\Sigma}}\exp(-\frac{1}{2\sigma^2}(\mathbf{y}-\mu)'\Sigma^{-1}(\mathbf{y}-\mu)) \\ &= k\cdot\sigma^{-n}\exp(-\frac{1}{2\sigma^2}\|\mathbf{y}-\mu\|^2). \end{aligned}$$

Med disse betegnelser har vi

SÆTNING 3.7. Lad situationen være som ovenfor. Da er kvotienttestet på niveau α af

$$H_0 : \mu \in H \quad \text{mod} \quad H_1 : \mu \in M\backslash H,$$

ækvivalent med testet givet ved det kritiske område

$$C_\alpha = \{(y_1,\ldots,y_n) | \frac{\|p_M(\mathbf{y})-p_H(\mathbf{y})\|^2/(k-r)}{\|\mathbf{y}-p_M(\mathbf{y})\|^2/(n-k)} > \mathrm{F}(k-r,n-k)_{1-\alpha}\}.$$

▲

BEVIS 3.6. Kvotienttesteststørrelsen er

$$\begin{aligned} Q &= \frac{\sup_{H_0}\mathrm{L}(\mu,\sigma^2)}{\sup\mathrm{L}(\mu,\sigma^2)} = \frac{\mathrm{L}(p_H(\mathbf{y}),\hat{\sigma^2})}{\mathrm{L}(p_M(\mathbf{y}),\hat{\sigma^2})} \\ &= \left[\frac{\|\mathbf{y}-p_M(\mathbf{y})\|^2}{\|\mathbf{y}-p_H(\mathbf{y})\|^2}\right]^{\frac{n}{2}}\frac{\exp(-\frac{n}{2})}{\exp(-\frac{n}{2})} = \left[\frac{\|\mathbf{y}-p_M(\mathbf{y})\|^2}{\|\mathbf{y}-p_H(\mathbf{y})\|^2}\right]^{\frac{n}{2}}. \end{aligned}$$

Heraf fås

$$Q < q \quad \Leftrightarrow \quad \frac{\|\mathbf{y} - p_M(\mathbf{y})\|^2}{\|\mathbf{y} - p_H(\mathbf{y})\|^2} < k_1.$$

Da vi forkaster hypotesen for små værdier af Q ses det, at vi netop forkaster, når længden af kateten $\mathbf{Y} - p_M(\mathbf{Y})$ er meget mindre end længden af hypotenusen. Da - ifølge Pythagoras' læresætning -

$$\|\mathbf{y} - p_H(\mathbf{y})\|^2 = \|\mathbf{y} - p_M(\mathbf{y})\|^2 + \|p_H(y) - p_M(y)\|^2,$$

ser vi, at man lige så godt kan sammenligne kateterne i.e. benytte,

$$Q < q \quad \Leftrightarrow \quad \frac{\|p_M(\mathbf{y}) - p_H(\mathbf{y})\|^2/(k-r)}{\|\mathbf{y} - p_M(\mathbf{y})\|^2/(n-k)} > c. \tag{3.1}$$

Under H_0 er såvel tæller som nævner $\sigma^2\chi^2(\mathrm{f})/\mathrm{f}$ fordelte med henholdsvis $k-r$ og $n-k$ frihedsgrader, og de er ydermere uafhængige (ifølge spaltningssætningen). Kvotienten vil derfor være F-fordelt under H_0, og sætningen følger.

Grunden til, at vi i 3.1 har divideret de pågældende normer med dimensionen af det relevante underrum er selvfølgelig, at vi ønsker at teststørrelsen skal være F-fordelt under H_0, og ikke blot proportional med en F-fordeling. ■

Man samler sædvanligt beregningerne i et variansanalyseskema.

Variation	SAK	Frihedsgrader = dimension
Af model fra hypotese	$\|p_M(\mathbf{Y}) - p_H(\mathbf{Y})\|^2$	$k-r$
Af observationer fra model	$\|\mathbf{Y} - p_M(\mathbf{Y})\|^2$	$n-k$
Af observationer fra hypotese	$\|\mathbf{Y} - p_H(\mathbf{Y})\|^2$	$n-r$

BEMÆRKNING 3.6. Hyppigt vil man være i den situation, at underrummene M og H er **parametriseret**, i.e.

$$\mu \in M \quad \Leftrightarrow \quad \exists\theta \in R^k(\mu = \mathbf{x}\theta)$$
$$\mu \in H \quad \Leftrightarrow \quad \exists\gamma \in R^r(\mu = \mathbf{x}_0\gamma),$$

hvor $\mathbf{x}$ og x_0 er $n \times k$ henholdsvis $n \times r$ $(r \leq k)$ matricer. Vi har da, at $p_M(\mathbf{y}) = \mathbf{x}\hat{\theta}$ og $p_H(\mathbf{y}) = \mathbf{x}_0\hat{\gamma}$ beregnes ved at løse ligningerne

$$\begin{aligned}(\mathbf{x}'\Sigma^{-1}\mathbf{x})\hat{\theta} &= \mathbf{x}'\Sigma^{-1}\mathbf{y} \\ (\mathbf{x}_0'\Sigma^{-1}\mathbf{x}_0)\hat{\gamma} &= \mathbf{x}_0'\Sigma^{-1}\mathbf{y}\end{aligned}$$

med hensyn til $\hat{\theta}$ og $\hat{\gamma}$. ▼

Vi betragter nu igen modellen fra p. 118.

EKSEMPEL 3.8. Vi har modellen

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} + \epsilon.$$

De observerede data var $y' = (10.11,\ 0.81,\ 5.24)$. Vi ønsker at teste hypotesen

$$H_0 : \theta_2 = 0 \quad \text{mod} \quad H_1 : \theta_2 \neq 0.$$

Vi omformulerer hypotesen til

$$H_0 : \mathrm{E}(\mathbf{Y}) = \begin{bmatrix} 1 \\ 0 \\ \frac{1}{2} \end{bmatrix} \theta_1 = \begin{bmatrix} 1 \\ 0 \\ \frac{1}{2} \end{bmatrix} \gamma.$$

Estimatoren for γ er

$$\hat{\gamma} = [(\ 1 \quad 0 \quad \tfrac{1}{2}\ ) \begin{bmatrix} 1 \\ 0 \\ \frac{1}{2} \end{bmatrix}]^{-1}[(\ 1 \quad 0 \quad \tfrac{1}{2}\ ) \begin{bmatrix} y1 \\ y2 \\ y3 \end{bmatrix}] = \frac{4}{5}y_1 + \frac{2}{5}y3.$$

Den observerede værdi er $\hat{\gamma} = 10.184$. Heraf fås

$$\mathbf{x}_0\hat{\gamma} = \begin{bmatrix} 1 \\ 0 \\ \frac{1}{2} \end{bmatrix} 10.184 = \begin{bmatrix} 10.184 \\ 0 \\ 5.092 \end{bmatrix},$$

og

$$\|\mathbf{y} - \mathbf{x}_0\hat{\gamma}\|^2 = (\mathbf{y} - \mathbf{x}_0\hat{\gamma})'(\mathbf{y} - \mathbf{x}_0\hat{\gamma}) = 0.6835.$$

Da vi havde (p. 119)

$$\|\mathbf{y} - \mathbf{x}_0\hat{\theta}\|^2 = (\mathbf{y} - \mathbf{x}\hat{\theta})'(\mathbf{y} - \mathbf{x}\hat{\theta}) = 0.0338,$$

får vi

$$\|\mathbf{x}\,\hat{\theta} - \mathbf{x}_0\hat{\gamma}\|^2 = 0.6835 - 0.0338 = 0.6497.$$

Følgelig er teststørrelsen

$$\frac{\|\mathbf{x}\,\hat{\theta} - \mathbf{x}_0\hat{\gamma}\|^2/(2-1)}{\|\mathbf{y} - \mathbf{x}\,\hat{\theta}\|^2/(3-2)} = 19.22 < \mathrm{F}(1,1)_{0.90},$$

og vi accepterer altså hypotesen, ihvert fald for $\alpha < 10\%$.
Forklaring på frihedsgraderne:

$$\mathrm{rg}\,\mathbf{x} = \mathrm{rg}\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} = 2 = k$$

$$\mathrm{rg}\,\mathbf{x}_0 = \mathrm{rg}\begin{bmatrix} 1 \\ 0 \\ \frac{1}{2} \end{bmatrix} = 1 = r$$

$$n = 3.$$

♦

Vi betragter dernæst en fortsættelse til eksempel 3.5 p. 125.

EKSEMPEL 3.9. Det forekommer ud fra problemstillingen rimeligt at antage, at parameteren $\theta_{21} = 0$. Vi vil derfor teste hypotesen

$$H_0 : \theta_{21} = 0 \quad \text{mod} \quad H_1 : \theta_{21} \neq 0.$$

Hypoteserummet H er derfor givet ved, at

$$\mathrm{E}(\mathbf{Y}) = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \theta_{11} \\ \theta_{12} \end{bmatrix} = \begin{bmatrix} \mu_1 + \theta_{11} \\ \mu_1 + \theta_{11} \\ \mu_1 + \theta_{12} \\ \mu_1 + \theta_{12} \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Vi finder nu

$$\mathbf{x}_1'\mathbf{x}_1 = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 2 & 0 \\ 2 & 0 & 2 \end{bmatrix},$$

og

$$\mathbf{x}_1'\mathbf{Y} = \begin{bmatrix} Y_{1..} \\ Y_{11.} \\ Y_{12.} \end{bmatrix}.$$

Vi ser at $\mathbf{x}_1'\mathbf{x}_1$ er singulær, og vi indfører det lineære bånd

$$\mathbf{b}\,\theta = \begin{pmatrix} 0 & 1 & 1 \end{pmatrix} \begin{bmatrix} \mu_1 \\ \theta_{11} \\ \theta_{12} \end{bmatrix} = \theta_{11} + \theta_{12} = 0.$$

Da

$$\mathbf{b}'\mathbf{b} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix},$$

fås

$$\mathbf{x}'\mathbf{x} + \mathbf{b}'\mathbf{b} = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 3 & 1 \\ 2 & 1 & 3 \end{bmatrix}.$$

Denne matrix er inverteret p. 124. Vi finder derfor estimatoren under H_0 til

$$\hat{\theta}_1 = \begin{bmatrix} \frac{1}{2} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{2} & 0 \\ -\frac{1}{4} & 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} Y_{1..} \\ Y_{11.} \\ Y_{12.} \end{bmatrix} = \begin{bmatrix} \bar{Y}_{1..} \\ \bar{Y}_{11.} - \bar{Y}_1 \\ \bar{Y}_{12.} - \bar{Y}_1 \end{bmatrix}.$$

Den observerede værdi er $(-9,\ +8,\ -8)'$. Den nye residualvektor er

$$\mathbf{y} - \mathbf{x}_1\hat{\theta}_1 = (1, -1, -2. + 2, -6, 0, -2)'.$$

Normen af denne vektor er 50, og antallet af frihedsgrader er $7 - 2 = 5$. Vi finder derfor, at

$$\begin{aligned} \|p_M(\mathrm{y}) - p_H(y)\|^2 &= \|y - p_H(y)\|^2 - \|y - p_M(y)\|^2 \\ &= 50 - 28\frac{2}{3} = 21\frac{1}{3}. \end{aligned}$$

Vi kan nu samle beregningerne i nedenstående variansanalyseskema.

Variation	SAK	f	S^2	Test
$M - H$	$21\frac{1}{3}$	$3 - 2 = 1$	$21\frac{1}{3}$	2.97
$O - M$	$28\frac{2}{3}$	$7 - 3 = 4$	$7\frac{1}{6}$	
$O - H$	50	$7 - 2 = 5$		

Da den observerede værdi af teststørrelsen $2.97 < \mathrm{F}(1,4)_{0.90}$ vil vi acceptere hypotesen, og vi vil derfor gå ud fra, at H_0 er sand. ♦

3.2.2 Successiv testning i den generelle lineære model

Vi vil i dette afsnit illustrere fremgangsmåden ved den testprocedure man bør følge, når man **successivt ønsker at undersøge om middelværdivektoren for ens observationer ligger i underrum H_i med**

$$H_0 \supseteq H_1 \supseteq H_2 \supseteq \cdots \supseteq H_m, \qquad m \leq k.$$

Som udgangspunkt vælger vi at betragte følgende tal over udbyttet ved penicillingæring under anvendelse af 2 forskellige sukkerarter laktose og rørsukker ved koncentrationerne 2%, 4%, 6% og 8%. (i g./100ml.)

		Faktor B: koncentration			
		2%	4%	6%	8%
Faktor A:	Laktose	0.606	0.660	0.984	0.908
	Rørsukker	0.761	0.933	1.072	0.979

Tallene stammer fra [9] p. 314. Udbyttet er udtrykt ved logaritmen til mycelievægten efter en uges vækst.

Vi er nu interesserede i at undersøge de to faktorer A og B's indflydelse på det færdige resultat (udbytte). Vi antager, at observationerne er stokastisk uafhængige og normalt fordelte. Kaldes de

$$L : Y_{11}, Y_{12}, Y_{13}, Y_{14}$$

og

$$R : Y_{21}, Y_{22}, Y_{23}, Y_{24}$$

vil vi endvidere antage, at

$$\mathrm{E}(Y_{ij}) = \alpha_i' + \beta_i' x_j' + \gamma_i' x_j^2$$

hvor x_j' angiver j'te sukkerkoncentration. Vi foretager en skalaændring af sukkerkoncentrationen

2%	-3
4%	-1
6%	1
8%	3,

eller - mere stringent - definerer x ved

$$x_j = \frac{x_j' - 5\%}{1\%}.$$

Vi får da følgende udtryk for middelværdierne

$$\mathrm{E}(Y_{ij}) = \alpha_i + \beta_i x_j + \gamma_i x_j^2.$$

Vi antager altså, at udbyttet inden for de givne grænser kan beskrives ved andengradspolynomier.

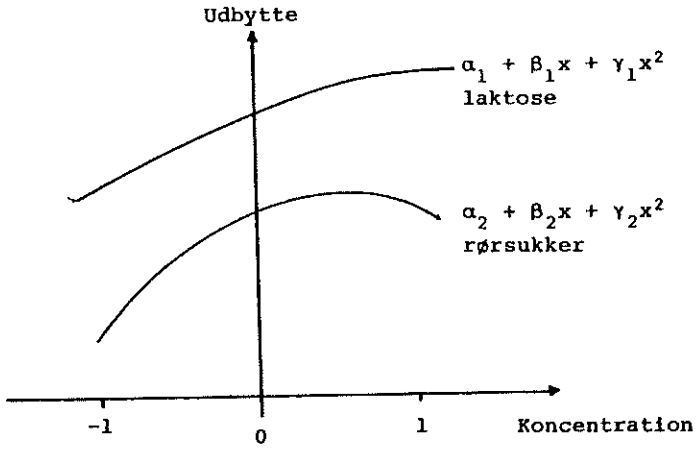
Man kan nu f.eks. succesivt undersøge

1) om $\gamma_1 = \gamma_2 = 0$, d.v.s. om en beskrivelse ved affine funktioner er tilstrækkelig,

2) - hvis dette accepteres - om $\beta_1 = \beta_2 = \beta$, d.v.s. om marginaleffekten ved at øge koncentrationen er den samme for de to sukkerarter,

3) - hvis dette accepteres - om $\alpha_1 = \alpha_2 = \alpha$, d.v.s. om de to sukkerarter er ens hvad angår udbyttet, og hvis dette accepteres,

4) om $\beta = 0$, d.v.s. om koncentrationen overhovedet har nogen indflydelse.



i) Vi skriver først modellen op på matrixform

$$\begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{13} \\ Y_{14} \\ Y_{21} \\ Y_{22} \\ Y_{23} \\ Y_{24} \end{bmatrix} = \begin{bmatrix} 1 & -3 & 9 & 0 & 0 & 0 \\ 1 & -1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 3 & 9 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -3 & 9 \\ 0 & 0 & 0 & 1 & -1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 3 & 9 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \beta_1 \\ \gamma_1 \\ \alpha_2 \\ \beta_2 \\ \gamma_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \end{bmatrix},$$

eller

$$\mathbf{Y} = \mathbf{x}\theta + \varepsilon.$$

Vi finder

$$\mathbf{x}'\mathbf{x} = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ -3 & -1 & 1 & 3 & 0 & 0 & 0 & 0 \\ 9 & 1 & 1 & 9 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & -3 & -1 & 1 & 3 \\ 0 & 0 & 0 & 0 & 9 & 1 & 1 & 9 \end{bmatrix} \begin{bmatrix} 1 & -3 & 9 & 0 & 0 & 0 \\ 1 & -1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 3 & 9 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -3 & 9 \\ 0 & 0 & 0 & 1 & -1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 3 & 9 \end{bmatrix}$$

$$= \begin{bmatrix} 4 & 0 & 20 & 0 & 0 & 0 \\ 0 & 20 & 0 & 0 & 0 & 0 \\ 20 & 0 & 164 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4 & 0 & 20 \\ 0 & 0 & 0 & 0 & 20 & 0 \\ 0 & 0 & 0 & 20 & 0 & 164 \end{bmatrix}.$$

Da

$$\begin{bmatrix} 4 & 0 & 20 \\ 0 & 20 & 0 \\ 20 & 0 & 164 \end{bmatrix}^{-1} = \begin{bmatrix} \frac{41}{64} & 0 & -\frac{5}{64} \\ 0 & \frac{1}{20} & 0 \\ -\frac{5}{64} & 0 & \frac{1}{64} \end{bmatrix},$$

er

$$(\mathbf{x}'\mathbf{x})^{-1} = \begin{bmatrix} \frac{41}{64} & 0 & -\frac{5}{64} & 0 & 0 & 0 \\ 0 & \frac{1}{20} & 0 & 0 & 0 & 0 \\ -\frac{5}{64} & 0 & \frac{1}{64} & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{41}{64} & 0 & -\frac{5}{64} \\ 0 & 0 & 0 & 0 & \frac{1}{20} & 0 \\ 0 & 0 & 0 & -\frac{5}{64} & 0 & \frac{1}{64} \end{bmatrix}.$$

Følgelig er

$$\hat{\theta} = \begin{bmatrix} -\frac{1}{16}y_{11} + \frac{9}{16}y_{12} + \frac{9}{16}y_{13} - \frac{1}{16}y_{14} \\ -\frac{3}{20}y_{11} - \frac{1}{20}y_{12} + \frac{1}{20}y_{13} + \frac{3}{20}y_{14} \\ \frac{1}{16}y_{11} - \frac{1}{16}y_{12} - \frac{1}{16}y_{13} + \frac{1}{16}y_{14} \\ -\frac{1}{16}y_{21} + \frac{9}{16}y_{22} + \frac{9}{16}y_{23} - \frac{1}{16}y_{24} \\ \frac{3}{20}y_{21} - \frac{1}{20}y_{22} - \frac{1}{20}y_{23} + \frac{3}{20}y_{24} \\ \frac{1}{16}y_{21} - \frac{1}{16}y_{22} - \frac{1}{16}y_{23} + \frac{1}{16}y_{24} \end{bmatrix} = \begin{bmatrix} 0.830 \\ 0.062 \\ -0.008 \\ 1.019 \\ 0.040 \\ -0.017 \end{bmatrix}.$$

Modellen svarer til et 6-dimensionalt underrum M i R^8 ($\mathrm{rg}\,\mathbf{x} = 6$), og vi har - idet vi regner med normen svarende til $\mathbf{I}$ - at projektionen på M er

$$p_M(\mathbf{y}) = x\hat{\theta} = \begin{bmatrix} 1 & -3 & 9 & 0 & 0 & 0 \\ 1 & -1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 3 & 9 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -3 & 9 \\ 0 & 0 & 0 & 1 & -1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 3 & 9 \end{bmatrix} \begin{bmatrix} 0.830 \\ 0.062 \\ -0.008 \\ 1.019 \\ 0.040 \\ -0.017 \end{bmatrix} = \begin{bmatrix} 0.572 \\ 0.760 \\ 0.884 \\ 0.944 \\ 0.746 \\ 0.962 \\ 1.042 \\ 0.986 \end{bmatrix}.$$

Vi har derfor residualerne

$$\mathbf{y} - p_M(\mathbf{y}) = \begin{bmatrix} 0.034 \\ -0.100 \\ 0.100 \\ -0.036 \\ 0.015 \\ -0.029 \\ 0.030 \\ -0.007 \end{bmatrix}.$$

Den kvadrerede længde af denne vektor er

$$\|\mathbf{y} - p_M(\mathbf{y})\|^2 = 0.034^2 + \cdots + (-0.007)^2 = 0.024467.$$

Som skøn over σ^2 kan derfor anvendes

$$\hat{\sigma}^2 = \frac{1}{8-6} 0.024467 = 0.0122335.$$

ii) Hvis hypotesen $\mu \in H_1$, d.v.s. $\gamma_1 = \gamma_2 = 0$, eller

$$\mathbf{y} = \begin{bmatrix} 1 & -3 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 3 & 0 & 0 \\ 0 & 0 & 1 & -3 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 3 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \beta_1 \\ \alpha_2 \\ \beta_2 \end{bmatrix} + \varepsilon = \mathbf{x}_1 \delta_1 + \varepsilon_1,$$

er sand, får vi estimaterne

$$\hat{\delta}_1 = (\mathbf{x}_1'\mathbf{x}_1)^{-1}\mathbf{x}_1'\mathbf{y} = \begin{bmatrix} \frac{1}{4}y_{11} + \frac{1}{4}y_{12} + \frac{1}{4}y_{13} + \frac{1}{4}y_{14} \\ -\frac{3}{20}y_{11} - \frac{1}{20}y_{12} + \frac{1}{20}y_{13} + \frac{3}{20}y_{14} \\ \frac{1}{4}y_{21} + \frac{1}{4}y_{22} + \frac{1}{4}y_{23} + \frac{1}{4}y_{24} \\ -\frac{3}{20}y_{21} - \frac{1}{20}y_{22} + \frac{1}{20}y_{23} + \frac{3}{20}y_{24} \end{bmatrix} = \begin{bmatrix} 0.790 \\ 0.062 \\ 0.936 \\ 0.040 \end{bmatrix}$$

Residualerne bliver

$$\mathbf{y} - p_{H_1}(\mathbf{y}) = \mathbf{y} - \mathbf{x}_1\hat{\delta}_1 = \begin{bmatrix} 0.002 \\ -0.068 \\ 0.132 \\ -0.068 \\ -0.055 \\ 0.037 \\ 0.096 \\ -0.077 \end{bmatrix}.$$

Den kvadrerede længde af denne vektor er

$$\|\mathbf{y} - p_{H_1}(\mathbf{y})\|^2 = 0.002^2 + \cdots + (-0.077)^2 = 0.046215.$$

iii) Hvis $\mu \in H_2$, d.v.s. $\beta_1 = \beta_2 = \beta$, er modellen

$$\mathbf{y} = \begin{bmatrix} 1 & 0 & -3 \\ 1 & 0 & -1 \\ 1 & 0 & 1 \\ 1 & 0 & 3 \\ 0 & 1 & -3 \\ 0 & 1 & -1 \\ 0 & 1 & 1 \\ 0 & 1 & 3 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \beta \end{bmatrix} + \varepsilon_2 = x_2\delta_2 + \varepsilon_2.$$

Estimaterne bliver

$$\hat{\delta}_2 = (\mathbf{x}_2'\mathbf{x}_2)^{-1}\mathbf{x}_2'\mathbf{y} = \begin{bmatrix} 0.790 \\ 0.936 \\ 0.051 \end{bmatrix},$$

og residualerne

$$\mathbf{y} - p_{H_2}(\mathbf{y}) = \begin{bmatrix} -0.031 \\ -0.079 \\ 0.143 \\ -0.035 \\ -0.022 \\ 0.048 \\ 0.085 \\ -0.110 \end{bmatrix}.$$

Den kvadrerede norm af residualvektoren er

$$||\mathbf{y} - p_{H_2}(\mathbf{y})||^2 = (-0.031)^2 + \cdots + (-0.110)^2 = 0.050989.$$

iv) Hvis $\mu \in H_3$, d.v.s. $\beta_1 = \beta_2 = \beta$ og $\alpha_1 = \alpha_2 = \alpha$, er modellen

$$\mathbf{y} = \begin{bmatrix} 1 & -3 \\ 1 & -1 \\ 1 & 1 \\ 1 & 3 \\ 1 & -3 \\ 1 & -1 \\ 1 & 1 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \varepsilon_3 = \mathbf{x}_3\delta_3 + \varepsilon_3$$

Vi finder

$$\hat{\delta}_3 = (\mathbf{x}_3'\mathbf{x}_3)^{-1}\mathbf{x}_3'\mathbf{y} = \begin{bmatrix} 0.863 \\ 0.051 \end{bmatrix},$$

og

$$\mathbf{y} - p_{H_3}(\mathbf{y}) = \begin{bmatrix} -0.104 \\ -0.152 \\ 0.070 \\ -0.108 \\ 0.051 \\ 0.121 \\ 0.158 \\ -0.037 \end{bmatrix},$$

hvorfor

$$\|\mathbf{y} - p_{H_3}(\mathbf{y})\|^2 = 0.094059.$$

v) Endelig betragtes tilfældet $\mu \in H_4$, d.v.s. $\beta = 0$, eller

$$\mathbf{y} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \alpha = \mathbf{x}_4 \delta_4 + \varepsilon_4.$$

Vi finder da

$$\hat{\delta}_4 = \hat{\alpha} = (\mathbf{x}_4'\mathbf{x}_4)^{-1}\mathbf{x}_4'\mathbf{y}' = 0.863,$$

hvorfor

$$\mathbf{y} - p_{H_4}(\mathbf{y}) = \begin{bmatrix} -0.250 \\ -0.203 \\ 0.121 \\ 0.045 \\ -0.102 \\ 0.070 \\ 0.209 \\ 0.116 \end{bmatrix},$$

og

$$\|\mathbf{y} - p_{H_4}(\mathbf{y})\|^2 = 0.196365.$$

Idet vi sætter $\mathrm{rg}(\mathbf{x}_i) = r_i$ og $\mathrm{rg}(\mathbf{x}) = k$ kan vi sammenfatte testningen i et variansanalyseskema som

Variation	SAK	Frihedsgrader = dimension
$H_4 - H_3$	$\|p_{H_4}(\mathbf{y}) - p_{H_3}(\mathbf{y})\|^2$	$r_3 - r_4 = 2 - 1 = 1$
$H_3 - H_2$	$\|p_{H_3}(\mathbf{y}) - p_{H_2}(\mathbf{y})\|^2$	$r_2 - r_3 = 3 - 2 = 1$
$H_2 - H_1$	$\|p_{H_2}(\mathbf{y}) - p_{H_1}(\mathbf{y})\|^2$	$r_1 - r_2 = 4 - 3 = 1$
$H_1 - M$	$\|p_{H_1}(\mathbf{y}) - p_M(\mathbf{y})\|^2$	$k - r_1 = 6 - 4 = 2$
M – obs.	$\|p_M(\mathbf{y}) - \mathbf{y}\|^2$	$n - k = 8 - 6 = 2$
H_4 – obs.	$\|p_{H_4}(\mathbf{y}) - \mathbf{y}\|^2$	$n - r_4 = 8 - 1 = 7$

Dette skema er en simpel udvidelse af skemaet p. 143. Vi kan anvende spaltningssætningen og får - under de forskellige hypoteser - at SAK'erne er uafhængige og er fordelt som $\sigma^2\chi^2$ med de anførte frihedsgrader.

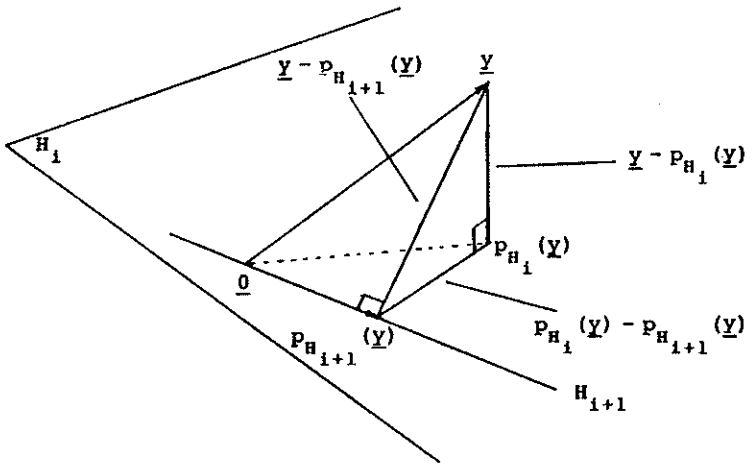
Hvis en hypotese H_i er accepteret bliver teststørrelsen for test af H_{i+1}

$$\frac{\|p_{H_i}(\mathbf{y}) - p_{h_{i+1}}(\mathbf{y})\|^2/(r_i - r_{i+1})}{\|p_{H_i}(\mathbf{y}) - \mathbf{y}\|^2/(n - r_i)}.$$

Under hypotesen er denne størrelse $\mathrm{F}(r_i - r_{i+1}, n - r)$ - fordelt (ifølge spaltningssætningen), og vi forkaster - stadig ifølge teorien fra forrige afsnit - for store værdier af Z, d.v.s. for

$$Z > \mathrm{F}(r_i - r_{i+1}, n - r_i)_{1-\alpha}.$$

Inden vi konkret går i gang med testningen, vil det være på sin plads at anføre nogle beregningsformler. Vi betragter overgangen fra H_i til $H_{i+1} \subset H_i$.



Ved hjælp af Pythagoras sætning ser vi nu, at der findes to alternative beregningsmåder for

$$z = \|p_{H_{i+1}}(\mathbf{y}) - p_{H_i}(\mathbf{y})\|^2,$$

nemlig

$$z = \|p_{H_i}(\mathbf{y})\|^2 - \|p_{H_{i+1}}(\mathbf{y})\|^2 \tag{3.2}$$

og

$$z = \|\mathbf{y} - p_{H_{i+1}}(\mathbf{y})\|^2 - \|\mathbf{y} - p_{H_i}(\mathbf{y})\|^2. \tag{3.3}$$

Af disse må **den første foretrækkes af numeriske grunde**, men hvis man alligevel har beregnet residualerne kvadratsummer er det åbenbart lettest at bruge (ii).

Variansanalyseskemaet bliver i vores tilfælde

Variation	SAK	f	Teststørrelse
$H_4 - H_3$	0.102306	1	$\frac{0.102306/1}{0.094059/6} = 5.44$
$H_3 - H_2$	0.043070	1	$\frac{0.043070/1}{0.050981/5} = 4.22$
$H_2 - H_1$	0.004774	1	$\frac{0.004774/1}{0.046215/4} = 0.41$
$H_1 - M$	0.021748	2	$\frac{0.021748/1}{0.024467/2} = 0.89$
M – obs	0.024467	2	
Obs – H_4	0.196365	7	

Da

$$4.22 \simeq \mathrm{F}(1,5)_{0.91},$$

og

$$5.44 \simeq \mathrm{F}(1,6)_{0.94},$$

vil man ikke - ved test på e.g. niveau $\alpha = 5\%$ - afvise nogen af hypoteserne H_1, H_2, H_3 eller H_4.

NOTE 1. Bemærk at vi selvfølgelig ikke ville teste e.g. H_2, hvis vi havde fået forkastet H_1, idet H_2 jo er en delhypotese af H_1.

Konklusionen må derfor bliver, at vi - indtil evt. nye undersøgelser afkræfter dette - vil arbejde med den model, at udbyttet Y ved penicillingæringen er uafhængig af sukkerart og den koncentration ($2\% \leq$ konc. $\leq 8\%$) ved hvilken gæringen sker. Vi har med

$$\mathrm{E}(Y) = \alpha \quad \text{og} \quad \mathrm{V}(Y) = \sigma^2,$$

at

$$\hat{\alpha} = 0.863,$$

og

$$\hat{\sigma}^2 = \frac{0.196365}{7} = 0.028052 \simeq 0.17^2.$$

Endvidere er

$$\mathrm{V}(\hat{\alpha}) = \frac{\sigma^2}{8} \simeq \frac{\hat{\sigma}^2}{8} = 0.0035 \simeq 0.059^2.$$

Kapitel 4
Regressionsanalyse

I dette kapitel giver vi en oversigt over regressionsanalysen. Det meste fremtræder som specialtilfælde af den generelle lineære model, men da en række anvendelser ofte er knyttet til regressionssituationer, vil vi beskrive resultaterne i dette sprog.

Der er anført et lille afsnit om ortogonal regression (ikke at forveksle med regression efter ortogonale polynomier). Rent statistisk hører dette sådan set hjemme under afsnittet om principale komponenter og faktoranalysen, og hvad angår beregningsmetoder, henvises også til dette afsnit. Jeg har dog ud fra et "curve-fitting" synspunkt fundet det formålstjenligt også at nævne begrebet her i dette kapitel.

4.1 Lineær regressionsanalyse

I dette afsnit vil den lineære regressionsanalyse blive analyseret v.h.a. teorien for den generelle lineære model. Vi indleder med

4.1.1 Notation og model

I den almindelige regressionsanalyse arbejder vi med modellen

$$\mathrm{E}(Y) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k,$$

hvor x'erne er kendte størrelser og β'erne (og α) er ukendte parametre. Hvis vi har givet n observationer af Y kan vi mere præcist skrive modellen

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{k1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{bmatrix} \cdot \begin{bmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

eller

$$\mathbf{Y} = \mathbf{x}\beta + \varepsilon.$$

Vi forudsætter som sædvanligt, at

$$\mathrm{D}(\varepsilon) = \sigma^2\boldsymbol{\Sigma},$$

hvor $\boldsymbol{\Sigma}$ er kendt og σ^2 (oftest) ukendt.

Estimatorerne findes på sædvanlig vis ved at løse **normalligningerne**

$$\mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{x}\beta = \mathbf{x}'\boldsymbol{\Sigma}^{-1}\mathbf{Y},$$

eller, hvis $\boldsymbol{\Sigma} = \mathbf{I}$

$$\mathbf{x}'\mathbf{x}\hat{\beta} = \mathbf{x}'\mathbf{Y}.$$

I det første tilfælde taler vi om en **vægtet regressionsanalyse**.

Inden vi går videre, vil det nok være hensigtsmæssigt endnu engang at præcisere, hvad der menes med ordet **lineær** i vendingen **lineær regressionsanalyse**.

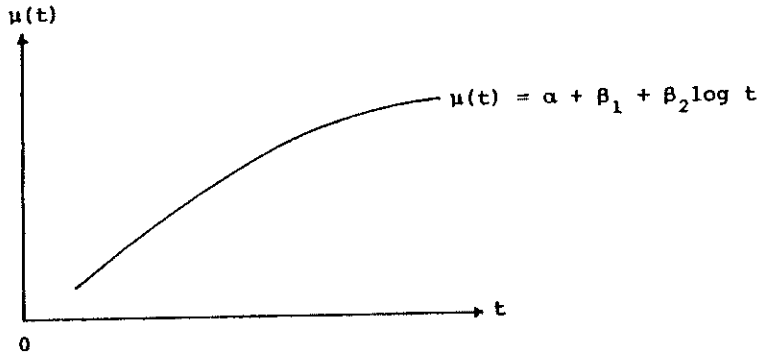
Meningen er - som i den almindelige lineære model - at der er tale om linearitet i **parametrene**. Vi kan sagtens lave regression efter e.g. tiden og logaritmen til tiden. Modellen er da blot

$$\mathrm{E}(Y) = \alpha + \beta_1 t + \beta_2 \log t,$$

jvf. eksempel 3.2

med n observationer bliver modellen på matrixform

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & t_1 & \log t_1 \\ \vdots & \vdots & \vdots \\ 1 & t_n & \log t_n \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$



Figur 4.1:

En anden banal ting, det kan være nyttigt at pointere, at at man kan **"tvinge" regressionsfladen gennem 0** ved at stryge α'et og 1'ste søjle i x-matricen, i.e. anvende modellen

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{k1} \\ \vdots & & \vdots \\ x_{1n} & \cdots & x_{kn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Det kan være nyttigt at bemærke, at man kan anvende følgende trick, når man ønsker, at regressionsfladen skal gå gennem 0. Vi forudsætter, at $\boldsymbol{\Sigma} = \mathbf{I}$.

Vi betragter observationerne $Y_1, \ldots, Y_n$ og de tilsvarende værdier af de uafhængige variable $1, x_{i1}, \ldots, x_{ik}$, $i = 1, \ldots, n$.
Hvis vi hertil føjer $-Y_1, \ldots, -Y_n$ og $1, -x_{i1}, \ldots, -x_{ik}$, $i = 1, \ldots, n$, og skriver den sædvanlige model op, fås

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \\ -Y_1 \\ \vdots \\ -Y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{k1} \\ \vdots & \vdots & & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \\ 1 & -x_{11} & \cdots & -x_{k1} \\ \vdots & \vdots & & \vdots \\ 1 & -x_{1n} & \cdots & -x_{kn} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \varepsilon,$$

eller mere kompakt skrevet

$$\begin{bmatrix} \mathbf{Y} \\ -\mathbf{Y} \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \mathbf{x} \\ \mathbf{1} & -\mathbf{x} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \varepsilon,$$

hvor vi har anvendt en i forhold til notationen p. 158 lidt afvigende definition af x-matricen og β.

Normalligningerne bliver

$$\begin{bmatrix} \mathbf{1}' & \mathbf{1}' \\ \mathbf{x}' & -\mathbf{x}' \end{bmatrix} \begin{bmatrix} \mathbf{1} & \mathbf{x} \\ \mathbf{1} & -\mathbf{x} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \mathbf{1}' & \mathbf{1}' \\ \mathbf{x}' & -\mathbf{x}' \end{bmatrix} \begin{bmatrix} \mathbf{Y} \\ -\mathbf{Y} \end{bmatrix},$$

eller

$$\begin{bmatrix} 2n & 0 \\ 0 & 2\mathbf{x}'\mathbf{x} \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} 0 \\ 2\mathbf{x}'\mathbf{Y} \end{bmatrix}.$$

Skrives disse ligninger ud, fås

$$\begin{aligned} 2n\alpha &= 0 \\ 2\mathbf{x}'\mathbf{x}\beta &= 2\mathbf{x}'\mathbf{Y}, \end{aligned}$$

eller

$$\begin{aligned} \alpha &= 0 \\ \mathbf{x}'\mathbf{x}\beta &= \mathbf{x}'\mathbf{Y}. \end{aligned}$$

Med andre ord får vi på denne måde bestemt estimatorerne til koefficienterne i en regressionsflade, der er tvunget gennem $\mathbf{0}$.

Grunden til, at ovenstående ofte er anvendeligt, er, at man i en mængde standardprogrammer **ikke** kan tvinge fladen gennem $\mathbf{0}$. Ved hjælp af ovenstående lille trick kan problemet så omgås.

Output fra et sådant program må dog tolkes med varsomhed, idet alle SAK'er er dobbelt så store, som de bør være. E.g. vil residualkvadratafvigelsessummen være beregnet som

$$\begin{aligned} &\left(\begin{bmatrix} \mathbf{Y} \\ -\mathbf{Y} \end{bmatrix} - \begin{bmatrix} \mathbf{x}\hat{\beta} \\ -\mathbf{x}\hat{\beta} \end{bmatrix}\right)' \left(\begin{bmatrix} \mathbf{Y} \\ -\mathbf{Y} \end{bmatrix} - \begin{bmatrix} \mathbf{x}\hat{\beta} \\ -\mathbf{x}\hat{\beta} \end{bmatrix}\right) \\ &= ([\mathbf{Y} - \mathbf{x}\hat{\beta}]', [-\mathbf{Y} + \mathbf{x}\hat{\beta}]') \begin{bmatrix} \mathbf{Y} - \mathbf{x}\hat{\beta} \\ -\mathbf{Y} + \mathbf{x}\hat{\beta} \end{bmatrix} \\ &= 2[\mathbf{Y} - \mathbf{x}\hat{\beta}]'[\mathbf{Y} - \mathbf{x}\hat{\beta}], \end{aligned}$$

i.e. det dobbelte af den korrekte residualkvadratafvigelsessum.

De angivne frihedsgrader vil heller ikke være korrekte. Man kan da selv opstille en almindelig lineær model og finde de korrekte frihedsgrader ved dimensionsbetragtninger.

4.1.2 Korrelation og regression

I sætning 2.3.2 p. 92 er anført et resultat, der kan benyttes ved et test for, om den multiple korrelationskoefficient mellem normalt fordelte variable er 0. Vi skal nu vise, at dette resultat har en sammenhæng med et test i en regressionsmodel.

Vi antager, at vi har den sædvanlige model p. 157, og vi antager at $\Sigma = \mathbf{I}$.

Vi kan nu uden videre anvende teorien fra kapitel 3 til at teste diverse hypoteser om parametrene $\alpha, \beta_1, \ldots, \beta_k$.

Ved **formelle** regninger kan vi estimere den multiple korrelationskoefficient mellem Y og $x_1, \ldots, x_k$ ved hjælp af de udtryk, som er anført i afsnit 2.3.2.

Det kan vises, at vi får

$$R^2 = \frac{\|\mathbf{Y} - p_0(\mathbf{Y})\|^2 - \|\mathbf{Y} - p_M(\mathbf{Y})\|^2}{\|\mathbf{Y} - p_0(\mathbf{Y})\|^2},$$

hvor

$$p_0(\mathbf{Y}) = \begin{bmatrix} \bar{Y} \\ \vdots \\ \bar{Y} \end{bmatrix} (= \mathbf{x} \cdot \hat{\beta}),$$

og

$$p_M(\mathbf{Y}) = \mathbf{x}\beta = \hat{\mathrm{E}}(\mathbf{Y}).$$

Disse resultater er ikke særligt overraskende. Vi erindrer, at den multiple korrelationskoefficient bl.a. kan fås frem ved at finde den linearkombination af $\mathbf{X}$, der minimaliserer variansen af $(Y - \alpha'\mathbf{X})$, og dette svarer jo præcis til at opskrive betingelsen for mindste kvadraters estimater.

Sætter vi

$$\mathrm{SAK}_{\mathrm{tot}} = \|\mathbf{Y} - p_0(\mathbf{Y})\|^2 = \sum_i (Y_i - \bar{Y})^2,$$

og

$$\mathrm{SAK}_{\mathrm{res}} = \|\mathbf{Y} - \mathbf{x}\hat{\beta}\|^2 = \sum_i (Y_i - \hat{\mathrm{E}}(Y_i))^2,$$

kan vi skrive

$$R^2 = \frac{\mathrm{SAK}_{\mathrm{tot}} - \mathrm{SAK}_{\mathrm{res}}}{\mathrm{SAK}_{\mathrm{tot}}},$$

d.v.s. den kvadrerede multiple korrelationskoefficient kan også her udtrykkes som den del af den totale variation i Y'erne, som vi har fået ved hjælp af de uafhængige variable.

En lignende refortolkning af de partielle korrelationer er naturligvis også mulig.

Vi ser i øvrigt, at hvis vi formelt opskriver det p. 92 anførte test for $\rho_{Y|x_1,\ldots,x_k} = 0$, får vi

$$\begin{aligned}\frac{R^2}{1-R^2}\frac{n-k-1}{k} &= \frac{\|\mathbf{Y}-p_0(\mathbf{Y})\|^2 - \|\mathbf{Y}-p_M(\mathbf{Y})\|^2}{\|\mathbf{Y}-p_M(\mathbf{Y})\|^2}\frac{n-k-1}{k} \\ &= \frac{\|p_M(\mathbf{Y})-p_0(\mathbf{Y})\|^2/k}{\|\mathbf{Y}-P_M(\mathbf{Y})\|^2/(n-k-1)} \\ &= \frac{(\mathrm{SAK_{tot}}-\mathrm{SAK_{res}})/k}{\mathrm{SAK_{res}}/(n-k-1)}\end{aligned}$$

Ifølge den almindelige teori (p. 142) er dette netop teststørrelsen for hypotesen

$$\begin{bmatrix}\alpha\\ \beta_1\\ \vdots\\ \beta_k\end{bmatrix} = \begin{bmatrix}\alpha\\ 0\\ \vdots\\ 0\end{bmatrix},$$

og teststørrelsens fordeling er en $\mathrm{F}(k, n-k-1)$-fordeling - præcis det samme, som vi fik p. 142.

Det er, hvad angår de numeriske forhold ved testningen, altså underordnet, om vi opfatter x'erne som **realiserede udfald** af en **k-dimensionalt normalt fordelt stokastisk variabel** eller som **faste, deterministiske størrelser**, vi selv kan bestemme.

Dette punkt kan derfor holdes uden for den diskussion af forudsætningerne, vi betragter i næste afsnit.

4.1.3 Analyse af forudsætninger

Hvis man for samhørende x-værdier

$$x_{1i}, \ldots, x_{pi}$$

har flere observationer af Y, er det muligt at foretage de sædvanlige tests for fordelingstype (histogram, fraktildiagram,χ^2-test etc.) og for homogenitet af varianser (Bartletts test m.fl.). Endvidere kan vi lave runtest for tilfældighed etc.etc.

Nu er situationen den, at vi kun meget sjældent har (mere end højst et par) gentagelser for de forskellige værdier af den uafhængige variabel. Derfor er det ikke muligt at lave

disse undersøgelser af forudsætninger. Man betragter i stedet **residualerne**

$$\mathrm{E}_i = Y_i - \hat{\mathrm{E}}(Y_i) = Y_i - \hat{\alpha} - \hat{\beta}_1 x_{1i} - \hat{\beta}_2 x_{2i} - \cdots - \hat{\beta}_k x_{ki}.$$

Disse vil, såfremt modellen holder, være **approximativt** uafhængige og $\mathrm{N}(0, \sigma^2)$-fordelt.

Hvis man afbilder residualerne på forskellige måder og derved får noget frem, der ikke ser ud til at være (eller kunne være) realiserede udfald af indbyrdes uafhængige $\mathrm{N}(0, \sigma^2)$-fordelte stokastiske variable, har vi fået en indikation for, at der er noget i vejen med modellen.

Man vil vel oftest starte med en sædvanlig fordelingsanalyse af residualerne, i.e. lave runtest, tegne histogram, fraktildiagram etc.

Dernæst kan man afbilde residualerne mod forskellige størrelser (tiden, uafhængige variable etc.). Vi giver følgende 4 skitser til illustration af hyppigt forekommende **residualplots**. Vi vil nu give en kort beskrivelse af hvad årsagen til plots af disse udseender kan være. Vi konstaterer først, at 1 altid må anses for acceptabelt (jvf. dog p. 165).

i) Plot af residualer mod **tiden**

2) Variansen vokser med tiden. Lav en vægtet analyse

3) Mangler led af formen $\beta \cdot$tid

4) Mangler led af formen $\beta_1 \cdot \text{tid} + \beta_2 \cdot \text{tid}^2$

ii) Plot af residualer mod $\hat{\mathrm{E}}(Y_i)$

2) Variansen vokser med $\mathrm{E}(Y_i)$. Lav en vægtet analyse eller transformer Y'erne (e.g. med logaritme eller lign.)

3) Manglende konstantled (regressionen muligvis fejlagtigt tvunget gennem 0). Fejl i analyse.

4) Dårlig model. Forsøg med transformation af Y'erne.

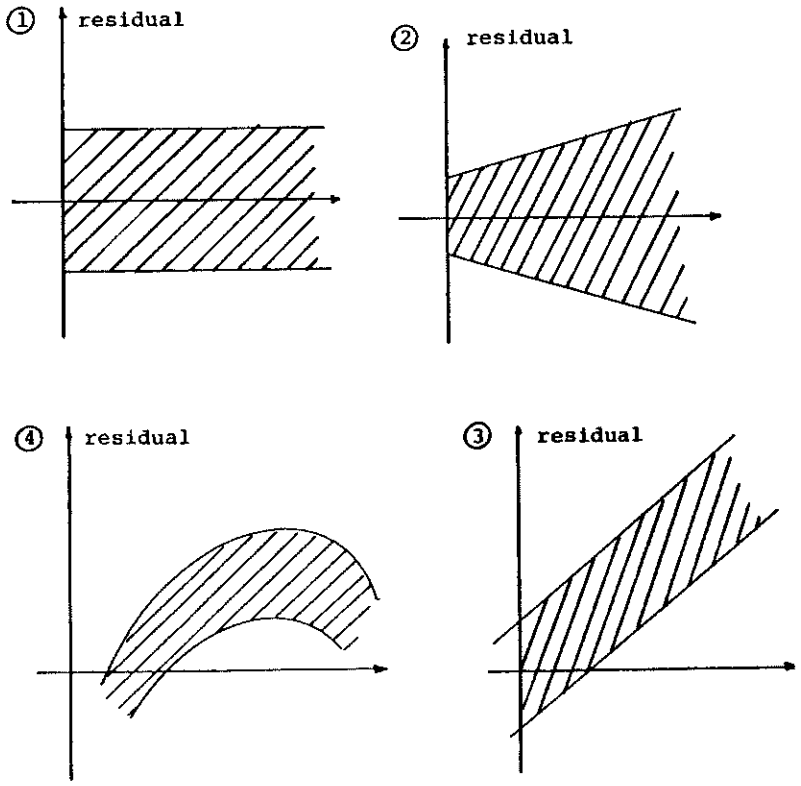
iii) Plot mod **uafhængig variabel** x_i

2) Variansen vokser med x_i. Lav en vægtet analyse eller transformer Y'erne

3) Fejl i beregningerne

4) Mangler kvadratisk led i x_i

Ovenstående er ikke tænkt som en udtømmende beskrivelse af hvorledes man analyserer residualplots, men nærmest som en vejledning i, efter hvilke retningslinier en sådan analyse kan foregå.



Figur 4.2: Residualplots.

BEMÆRKNING 4.1. Man ser ofte, at der laves residualplots af typen residual mod afhængig variabel, i.e.

$$Y_i - \hat{\mathrm{E}}(Y_i) \quad \text{mod} \quad Y_i,$$

Og der udtrykkes da ofte undren over, at billedet er som i 3). Deri ligger imidlertid intet abnormt. Det kan nemlig vises, at

$$\mathrm{Cor}(Y_i, Y_i - \hat{\mathrm{E}}(Y_i)) = 1 - R^2,$$

d.v.s. de er positivt **korrelerede**. Hvis den multiple korrelationskoefficient blot er lidt mindre end 1, vil man derfor få et billede som 3). Kun hvis regressionsfladen går **gennem samtlige** punkter, i.e. $R^2 = 1$, vil man få et billede som 1).

I praksis vil man ofte få sit residualplot printet ud på en linieskriver. Da kan plottene se ud som anført p. 166. De 4 plot er taget fra [32] p. 14-15 i appendix C.

Ved tolkningen af disse plot må man være opmærksom på, at der ikke altid er lige mange observationer for hver værdi af den uafhængige variable.

Dette er eksempelvis tilfældet ved det plot, der afbilder fejlen (residualet) mod variabel 10.

Der er 7 observationer svarende til $x_{10} \sim 0.2704\,\mathrm{E}04$ og 35 observationer svarende til $x_{10} \sim 0.7126\,\mathrm{E}03$. Variationsbredden for residualerne er nogenlunde den samme i de to tilfælde. **Hvis** residualerne svarende til de to værdier af x_{10} havde samme varians, ville man imidlertid forvente at variationsbredden for de mange observationer var den største.

Dette vil med andre ord sige, at hvis man har flest observationer omkring tyngdepunktet for en uafhængig variabel, skal et residualplot snarere være ellipseformet end af formen 1) for at være tilfredsstillende. ▼

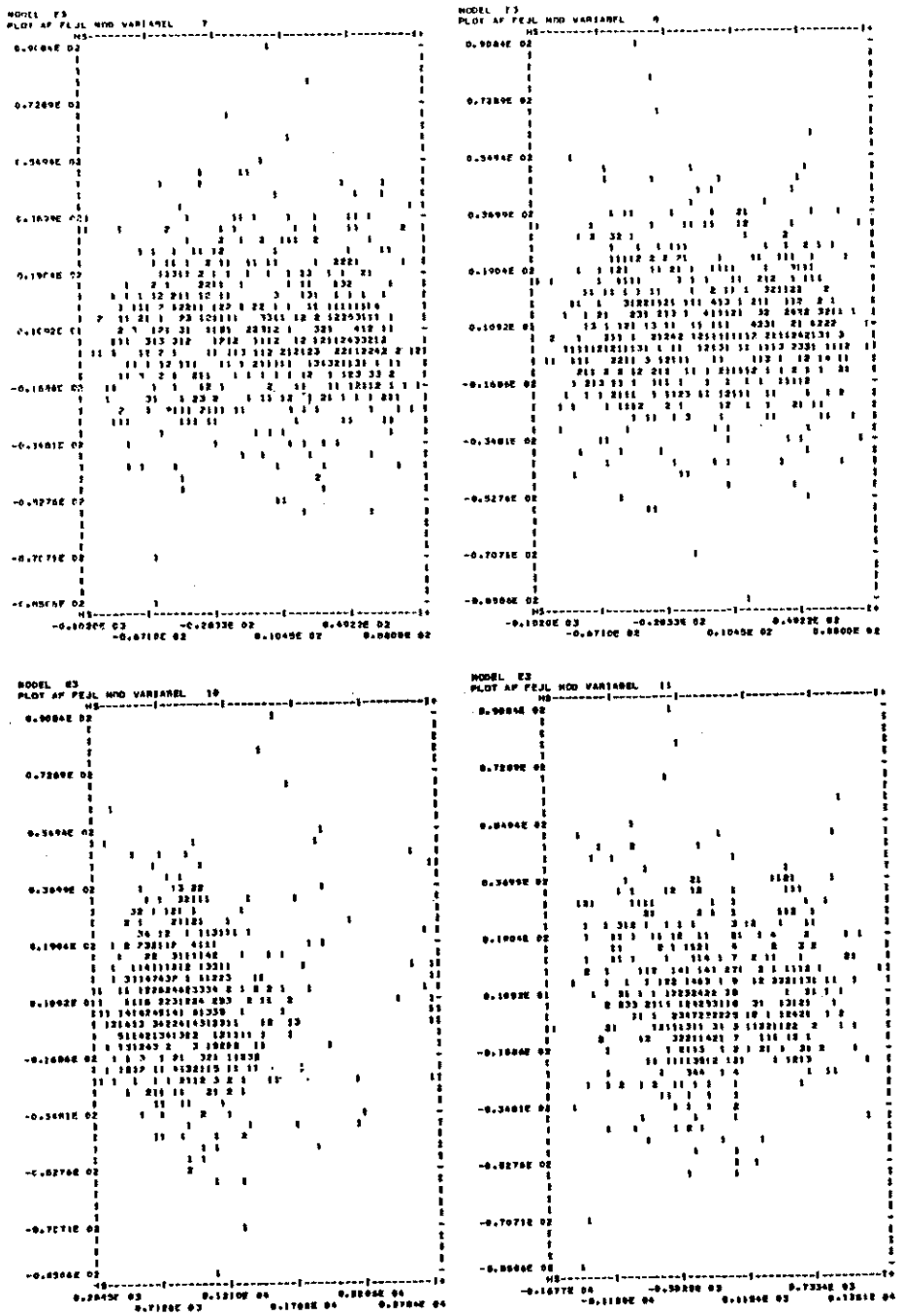
4.1.4 Noget om "Influence Statistics"

Ved vurderingen af kvaliteten af en regressionsanalyse går man ofte 2 veje

1) Undersøger om afvigelserne fra modellen synes tilfældige.

2) Undersøger effekten af enkeltmålinger på parameterestimater mv.

Betragtninger vedrørende 1) er givet i afsnit 4.1.3 ovenfor. Vi skal nedenfor kort redegøre for forhold omkring 2).



Figur 4.3:

Vi betragter modellen

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

dvs.

$$\mathbf{y} = \mathbf{x}\theta + \varepsilon$$

For den i'te række har vi

$$y_i = (x_{i1}, \cdots, x_{ip}) \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_p \end{pmatrix} + \varepsilon_i$$

eller

$$y_i = \mathbf{x}_i\theta + \varepsilon_i.$$

Vi antager, at $\varepsilon \in N(\mathbf{0}, \sigma^2\mathbf{I})$ og har derfor LS estimatet

$$\hat{\theta} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}$$

Den tilhørende residualvektor er

$$\mathbf{r} = \begin{pmatrix} r_1 \\ \vdots \\ r_n \end{pmatrix} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{x}\hat{\theta}$$

dvs.

$$\mathbf{r} = [\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}']\mathbf{y}$$

Dispersionsmatricerne for $\hat{\mathbf{y}}$ og $\mathbf{r}$ er

$$D(\hat{\mathbf{y}}) = \mathbf{x}D(\hat{\theta})\mathbf{x}' = \sigma^2\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'$$

$$\begin{aligned} D(\mathbf{r}) &= \sigma^2[\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'][\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'] \\ &= [\mathbf{I} + \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}' - 2\mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}]\sigma^2 \\ &= \sigma^2[\mathbf{I} - \mathbf{x}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}] \end{aligned}$$

For den i'te række finder vi

$$\begin{aligned} V(\hat{y}_i) &= \sigma^2 \mathbf{x}_i(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}_i' = \sigma^2 h_i \\ V(r_i) &= \sigma^2(1 - \mathbf{x}_i(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}_i') = \sigma^2(1-h_i) \end{aligned}$$

Deletion formlen

Genberegning af parameterestimater ved udelukkelse af en enkelt observation kan ske ved anvendelse af formlen

$$(\mathbf{A} - \mathbf{uv}')^{-1} = \mathbf{A}^{-1} + \frac{\mathbf{A}^{-1}\mathbf{uv}'\mathbf{A}^{-1}}{1 - \mathbf{v}'\mathbf{A}^{-1}\mathbf{u}},$$

hvor de involverede matricer forudsættes at existere. I tilfældet $\mathbf{A} = \mathbf{x}'\mathbf{x}$ og $\mathbf{u} = \mathbf{v} = \mathbf{x}_i'$ fås

$$(\mathbf{x}'\mathbf{x} - \mathbf{x}_i'\mathbf{x}_i)^{-1} = (\mathbf{x}'\mathbf{x})^{-1} + \frac{(\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}_i'\mathbf{x}_i(\mathbf{x}'\mathbf{x})^{-1}}{1-h_i}$$

Kaldes $\mathbf{x}$ -matricen med i'te række fjernet $\mathbf{x}(i)$ fås, at

$$\mathbf{x}(i)'\mathbf{x}(i) = \mathbf{x}'\mathbf{x} - \mathbf{x}_i'\mathbf{x}_i.$$

BEVIS 4.1. forbigås. ■

Vi er nu i stand til at opskrive de relevante udtryk.

Cook's D

Et konfidensområde for parameteren θ er alle de vektorer θ^*, der tilfredsstiller

$$\frac{1}{p\hat{\sigma}^2}(\hat{\theta} - \theta^*)'\mathbf{x}'\mathbf{x}(\hat{\theta} - \theta^*) \leq F(p, n-p)_{1-\alpha}.$$

Vi bruger venstresiden som et mål for en parametervektors afstand fra $\hat{\theta}$. Vi sætter $\hat{\theta}(i)$ lig det estimat, der fremkommer ved udelukkelse af den i'te observation,

$$\mathbf{y}(i) = (y_1, \cdots, y_{i-1}, y_{i+1}, \cdots, y_n)'$$

og har altså

$$\hat{\theta}(i) = [\mathbf{x}(i)'\mathbf{x}(i)]^{-1}\mathbf{x}(i)'\mathbf{y}(i).$$

Cooks D er da lig

$$\frac{1}{p\hat{\sigma}^2}(\hat{\theta}-\hat{\theta}(i))'\mathbf{x}'\mathbf{x}(\hat{\theta}-\hat{\theta}(i)).$$

Hvis Cook's D er lig fx. $F_{60\%}$ svarer dette altså til, at maximum likelihood skønnet flyttes ud på 60% konfidensellipsoiden for θ. Dette er en forholdsvis voldsom ændring blot ved fjernelse af en enkelt observation.

I SAS-programmet REG kan man finde Cooks D sammen med en række andre diagnostiske størrelser, som kort skal nævnes.

RSTUDENT & STUDENT RESIDUAL

RSTUDENT er et såkaldt "studentized" residual, dvs.

$$\text{RSTUDENT}_i = \frac{r_i}{\hat{\sigma}(i)\sqrt{1-h_i}},$$

hvor $\hat{\sigma}(i)^2$ er det variansskøn, der fremkommer ved udelukkelse af den i'te observation.

SAS beregner også en lignende størrelse, hvor den i'te observation ikke er udelukket

$$\text{STUDENT RESIDUAL} = \frac{r_i}{\hat{\sigma}\sqrt{1-h_i}}.$$

COVRATIO

COVRATIO måler ændringen i determinanten af dispersionsmatricen for parameterestimatet ved udelukkelse af den i'te måling. Vi finder

$$\text{COVRATIO}_i = \frac{\det[\hat{\sigma}(i)^2(\mathbf{x}(i)'\mathbf{x}(i))^{-1}]}{\det[\hat{\sigma}^2(\mathbf{x}'\mathbf{x})^{-1}]}$$

Denne størrelse "bør" ligge tæt på 1. Hvis den afviger meget, har den i'te observation for stor indflydelse.

DFFITS

DFFITS er ligesom Cooks distance et mål for den totale ændring forårsaget af udelukkelsen af en enkelt måling.

$$\begin{aligned}\text{DFFITS} &= \frac{\hat{y}_i-\hat{y}(i)_i}{\hat{\sigma}(i)\sqrt{h_i}} \\ &= \frac{\mathbf{x}_i[\hat{\theta}-\hat{\theta}(i)]}{\hat{\sigma}(i)\sqrt{h_i}}.\end{aligned}$$

DFBETAS

Hvor DFFITS måler ændringer i prediktionen af en observation forårsaget af ændringerne i samtlige parameterskøn, måler DFBETAS blot ændringen i et enkelt parameterskøn.

Vi har

$$\text{DFBETAS}_j = \frac{\hat{\theta}_j - \hat{\theta}(i)_j}{\hat{\sigma}(i)\sqrt{(\mathbf{x}'\mathbf{x})^{-1}_{jj}}}.$$

Kald i SAS

Alle de nævnte størrelser kan findes ved hjælp af et simpelt kald i SAS, fx.

```
proc reg data = sundhed;
model ilt = maxplus loebetid/influence;
```

Modelordrer etc. er de samme i REG som i GLM. De diagnostiske tests kommer ved ordren `/influence`.

4.2 Regression efter ortogonale polynomier

Når man skal foretage en regressionsanalyse efter polynomier kan man ofte opnå ganske væsentlige beregningsmæssige besparelser ved at indføre de såkaldte ortogonale polynomier. Dette giver i sidste instans samme udtryk for estimater af middelværdien som funktion af den uafhængige variabel, men altså med et væsentligt mindre regnearbejde.

4.2.1 Definition og modelformulering

Vi antager, at der er givet en polynomial regressionsmodel, i.e., at

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \xi_0(t_1) & \xi_1(t_1) & \cdots & \xi_k(t_1) \\ \vdots & \vdots & & \vdots \\ \xi_0(t_n) & \xi_1(t_n) & \cdots & \xi_k(t_n) \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Her betegner ξ_i, $i = 0, 1, \ldots, k$ kendte polynomier af i'te grad i t. Vi forudsætter, at

$$\begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \in \mathrm{N}(0, \sigma^2 \mathbf{I})$$

I denne model kan vi på helt sædvanlig måde estimere og teste hypoteser angående parametrene $(\alpha, \beta_1, \ldots, \beta_k)$.

Som anført indledningsvis kan man her med stor fordel betragte såkaldte ortogonale polynomier ξ_i, idet beregningsarbejdet da reduceres væsentligt.

Vi indfører disse polynomier i

DEFINITION 4.1. Ved et sæt **ortogonale polynomier** svarende til værdierne $t_1, \ldots, t_n$ forstås polynomier $\xi_0, \xi_1, \ldots$, hvor ξ_i er af i'te grad, som tilfredsstiller

$$\sum_{j=1}^{n} \xi_i(t_j) = 0, \qquad i = 1, 2, \ldots, k \tag{4.1}$$

$$\sum_{j=1}^{n} \xi_\mu(t_j)\xi_\gamma(t_j) = 0, \qquad \mu \neq \gamma. \tag{4.2}$$

▲

BEMÆRKNING 4.2. Det påpeges, at ξ_0 er en konstant, hvorfor 4.1 selvfølgelig ikke finder anvendelse på ξ_0. Af notationsmæssige grunde sættes $\xi_i(t_j) = \xi_{ij}, \forall_{i,j}$. Vi skal senere vende tilbage til problemet med bestemmelse af ortogonale polynomier. ▼

Forudsætter vi nu, at polynomierne i selve modellen er ortogonale, fås med

$$\xi = \begin{bmatrix} \xi_0 & \cdots & \xi_{k1} \\ \vdots & & \vdots \\ \xi_0 & \cdots & \xi_{kn} \end{bmatrix} = \begin{bmatrix} \xi_0(t_1) & \xi_1(t_1) & \cdots & \xi_k(t_1) \\ \vdots & \vdots & & \vdots \\ \xi_0(t_n) & \xi_1(t_n) & \cdots & \xi_k(t_n) \end{bmatrix},$$

at

$$\xi'\xi = \begin{bmatrix} n\xi_0^2 & 0 & \cdots & 0 \\ 0 & \sum \xi_{1j}^2 & & \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & & \sum \xi_{kj}^2 \end{bmatrix},$$

d.v.s. $\xi'\xi$ er en diagonalmatrix. Vi finder derfor

$$\hat{\beta} = (\xi'\xi)^{-1}\xi'\mathbf{Y} = \begin{bmatrix} \bar{Y}/\xi_0 \\ \sum \xi_{1j}Y_j / \sum \xi_{1j}^2 \\ \vdots \\ \sum \xi_{kj}Y_j / \sum \xi_{kj}^2 \end{bmatrix}$$

og

$$\mathrm{D}(\hat{\beta}) = \sigma^2 \begin{bmatrix} 1/n\xi_0^2 & 0 & \cdots & 0 \\ 0 & 1/\sum \xi_{1j}^2 & & \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & & 1/\sum \xi_{kj}^2 \end{bmatrix}.$$

Vi har derfor, at estimatorerne for parametrene er ukorrelerede, og da vi arbejder i en normal model altså også stokastisk uafhængige.

Vi finder, at residualkvadratafvigelsessummen er

$$\begin{aligned} \mathrm{SAK}_{\mathrm{res}} &= \|\mathbf{Y} - \xi\hat{\beta}\|^2 \\ &= (\mathbf{Y} - \xi\hat{\beta})'(\mathbf{Y} - \xi\hat{\beta}) \\ &= \mathbf{Y}'\mathbf{Y} - \hat{\beta}'\xi'\xi\hat{\beta} \\ &= \sum Y_j^2 - \{\hat{\alpha}^2 n\xi_0^2 + \hat{\beta}^2 \sum \xi_{1j}^2 + \cdots + \hat{\beta}_k^2 \sum \xi_{kj}^2\} \\ &= \sum (Y_j - \bar{Y})^2 - \{\hat{\beta}^2 \sum \xi_{1j}^2 + \cdots + \hat{\beta}_k^2 \sum \xi_{kj}^2.\} \end{aligned}$$

Heraf fås nu umiddelbart følgende

SÆTNING 4.1. Vi har følgende spaltning af den totale variation

$$\sum (Y_j - \bar{Y})^2 = \hat{\beta}_1^2 \sum \xi_{1j}^2 + \cdots + \hat{\beta}_k^2 \sum \xi_{kj}^2 + \sum \{Y_j - \bar{Y} - \hat{\beta}_1 \xi_1(t_j) - \cdots - \hat{\beta}_k \xi_k(t_j)\}^2,$$

eller - med en letfattelig notation -

$$\mathrm{SAK}_{\mathrm{tot}} = \mathrm{SAK}_{1.\mathrm{grad}} + \cdots + \mathrm{SAK}_{k.\mathrm{grad}} + \mathrm{SAK}_{\mathrm{res}},$$

d.v.s. den er spaltet i led svarende til hvert polynomium plus residualkvadratafvigelsessummen. Frihedsgraderne er $n-1$ henholdsvis $1, \ldots, 1$ og $n-k-1$. ▲

BEVIS 4.2. Triviel følge af ovenstående ■

Ved hjælp af spaltningssætningen fås ydermere

SÆTNING 4.2. De i foregående sætning anførte kvadratsummer er stokastisk uafhængige med forventede værdier.

$$\begin{aligned} \mathrm{E}(\mathrm{SAK}_{i.\mathrm{grad}}) &= \mathrm{E}(\hat{\beta}_i^2 \sum_j \xi_i(t_j)^2) \\ &= \sigma^2 + \beta_i^2 \sum_j \xi_i(t_j)^2, \qquad i = 1, \ldots, k. \end{aligned}$$

og

$$\mathrm{E}(\mathrm{SAK}_{\mathrm{res}}) = \mathrm{E}[\sum_j (Y_j - \bar{Y} - \cdots - \hat{\beta}_k \xi_k(t_j))^2] = (n-k-1)\sigma^2.$$

Endelig er

$$\frac{1}{\sigma^2}\mathrm{SAK}_{\mathrm{res}} \in \chi^2(n-k-1),$$

og - såfremt $\beta_i = 0$ -

$$\frac{1}{\sigma^2}\mathrm{SAK}_{i.\mathrm{grad}} \in \chi^2(1).$$

▲

BEVIS 4.3. Umiddelbart. ■

Sætningerne indeholder de nødvendige resultater for at kunne opstille tests for hypoteserne

$$H_{0i} : \beta_i = 0 \quad \text{mod} \quad H_{1i} : \beta_i \neq 0.$$

Vi samler resultaterne i et **variansanalyseskema**

Variation	SAK	f	E(SAK/f)
Lineær	$SAK_{1.\text{grad}}$	1	$\sigma^2 + \beta_1^2 \sum_j \xi_1(t_j)^2$
Kvadratisk	$SAK_{2.\text{grad}}$	1	$\sigma^2 + \beta_2^2 \sum_j \xi_2(t_j)^2$
Kubisk	$SAK_{3.\text{grad}}$	1	$\sigma^2 + \beta_3^2 \sum_j \xi_3(t_j)^2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
k'te ordens	$SAK_{k.\text{grad}}$	1	$\sigma^2 + \beta_k^2 \sum_j \xi_k(t_j)^2$
Residual	SAK_{res}	$n-k-1$	σ^2
Total	SAK_{tot}	$n-1$	

BEMÆRKNING 4.3. Den store fordel ved at anvende ortogonale polynomier i regressionsanalysen er, at man uden at ændre nogle af de tidligere beregninger kan indføre polynomier af $(p+1)$'te orden, $(p+2)$'den orden etc. Ved bestemmelse af ordenen for det beskrivende polynomium vil man som regel fortsætte en (estimation og) testning indtil 2 på hinanden følgende β_i'er er lig 0, idet bidrag der skyldes lige ordens led og ulige ordens led er væsensforskellige. Dette er dog naturligvis en regel, der skal følges med varsomhed. Hvis man e.g. har en på de fysiske forhold baseret formodning om, at led af 5'te orden er af betydning, vil man selvfølgelig ikke slutte analysen, selv om 3.die og 4.de ordens koefficienterne ikke afviger signifikant fra 0. ▼

4.2.2 Bestemmelse af ortogonale polynomier

Det ses umiddelbart, at multiplikation med en konstant ikke ændrer ved ortogonalitetsbetingelserne 4.1 og 4.2. Vi vælger derfor at sætte

$$\xi_0(t) = \xi_0 = 1.$$

Polynomiet af 1.ste grad er

$$\xi_1(t) = t + a,$$

idet vi uden videre kan sætte koefficienten til t lig 1. Af 4.1 fås

$$0 = \sum_{j=1}^{n} \xi_1(t_j) = \sum_{j=1}^{n} (t_j + a) = \sum_{j=1}^{n} t_j + na,$$

eller

$$a = -\frac{1}{n}\sum_{j=1}^{n} t_j = -\bar{t},$$

d.v.s.

$$\xi_1(t) = t - \bar{t}.$$

Vi kan dernæst bestemme ξ_2 som en linearkombination af 1, ξ_1 og ξ_1^2, d.v.s.:

$$\xi_2(t) = a_{02} + a_{12}(t - \bar{t}) + a_{22}(t - \bar{t})^2.$$

Af 4.1 fås

$$0 = \sum_{j=1}^{n} \xi_2(t_j) = na_{02} + a_{12}\sum_j (t_j - \bar{t}) + a_{22}\sum_j (t_j - \bar{t})^2$$
$$\frac{a_{02}}{a_{22}} = -\frac{1}{n}\sum_j (t_j - \bar{t})^2.$$

Af 4.2 fås

$$\begin{aligned} 0 &= \sum_{j=1}^{n} \xi_1(t_j)\xi_2(t_j) \\ &= a_{02}\sum_j (t_j - \bar{t}) + a_{12}\sum_j (t_j - \bar{t})^2 + a_{22}\sum_j (t_j - \bar{t})^3 \\ &= a_{12}\sum_j (t_j - \bar{t})^2 + a_{22}\sum_j (t_j - \hat{t})^3. \end{aligned}$$

Heraf fås

$$\frac{a_{12}}{a_{22}} = -\frac{\sum_j (t_j - \bar{t})^3}{\sum_j (t_j - \bar{t})^2}.$$

ξ_3, ξ_4 etc. bestemmes på ganske analog vis.

Beregningerne bliver særligt simple, hvis t_i'erne er **ækvidistante**. Da sætter vi

$$u_j = \frac{t_j - (t_1 - w)}{w},$$

hvor $w = t_2 - t_1 = t_{i+1} - t_i$. Vi har så

$$u_i = i, \qquad i = 1, \ldots, n.$$

Svarende til **værdierne** $\mathbf{1}, \ldots, \mathbf{n}$ har vi da polynomierne givet ved

$$\xi_0(t) = 1 \tag{4.3}$$

$$\xi_1(t) = t - \frac{n+1}{2} \tag{4.4}$$

$$\xi_{i+1}(t) = \xi_1(t)\xi_i(t) - \frac{i^2(n^2 - i^2)}{4(4i^2 - 1)}\xi_{i-1}(t). \tag{4.5}$$

I omstående tabel p. 177 har vi givet værdier af ortogonale polynomier $\xi_1, \ldots, \xi_k$, $k \leq 5$,i $t = 1, \ldots, n$ for $n = 1, \ldots, 8$.

For at undgå brudne tal og store værdier har man valgt at anføre polynomier, hvor koefficienten til højestegradsleddet er et tal λ, der fremgår af tabellen. Endvidere er der angivet størrelserne

$$D = \sum_{j=1}^{n} \xi_i(j)^2 = \sum_{j=1}^{n} \xi_{ij}^2.$$

Vi giver nu et illustrativt

EKSEMPEL 4.1. I nedenstående tabel er der anført sammenhørende værdier mellem reaktionstemperaturen og udbyttet af en proces (i en fast tid).

Temperatur	**Udbytte**
200°F	0.75 oz.
210°F	1.00 oz.
220°F	1.35 oz.
230°F	1.80 oz.
240°F	2.60 oz.
250°F	3.60 oz.
260°F	5.45 oz.

Vi vil forsøge at beskrive udbyttet som en funktion af temperaturen ved hjælp af et polynomium. Vi antager, at forudsætningerne for at foretage en regressionsanalyse er til stede.

Vi transformerer først temperaturerne τ_i, $i = 1, \ldots, 7$ v.h.a. følgende relation

$$t_i = \frac{\tau_i - (200 - 10)}{10} = \frac{\tau_i - 190}{10}$$

n	3		4			5				6					7					8				
t	ξ_1	ξ_2	ξ_1	ξ_2	ξ_3	ξ_1	ξ_2	ξ_3	ξ_4	ξ_1	ξ_2	ξ_3	ξ_4	ξ_5	ξ_1	ξ_2	ξ_3	ξ_4	ξ_5	ξ_1	ξ_2	ξ_3	ξ_4	ξ_5
1	−1	1	−3	1	−1	−2	2	−1	1	−5	5	−5	1	−1	−3	5	−1	3	−1	−7	7	−7	7	−7
2	0	−2	−1	−1	3	−1	−1	2	−4	−3	−1	7	−3	5	−2	0	1	−7	4	−5	1	5	−13	23
3	1	1	1	−1	−3	0	−2	0	6	−1	−4	4	3	−10	−1	−3	1	1	−5	−3	−3	7	−3	−17
			3	1	1	1	−1	−2	−4	1	−4	−4	2	10	0	−4	0	6	0	−1	−5	3	9	−15
						2	2	1	1	3	−1	−7	−3	−5	1	−3	−1	1	5	1	−5	−3	9	15
										5	5	5	1	1	2	0	−1	−7	−4	3	−3	−7	−3	17
															3	5	1	3	1	5	1	−5	−13	−23
																				7	7	7	7	7
D	2	6	20	4	20	10	14	10	70	70	84	180	28	252	28	84	6	154	84	168	168	264	616	2184
λ	1	3	2	1	$\frac{10}{3}$	1	1	$\frac{5}{6}$	$\frac{35}{12}$	2	$\frac{3}{2}$	$\frac{5}{3}$	$\frac{7}{12}$	$\frac{21}{10}$	1	1	$\frac{1}{6}$	$\frac{7}{12}$	$\frac{7}{20}$	2	1	$\frac{2}{3}$	$\frac{7}{12}$	$\frac{7}{10}$

Tabel 4.1: Værdier af ortogonale polynomier.

Vi får da værdierne $t_1, \ldots, t_7 = 1, \ldots, 7$.

Vi anfører beregningerne i følgende skema

t_j	ξ_1	ξ_2	ξ_3	ξ_4	ξ_5	y_j
1	-3	5	-1	3	-1	0.75
2	-2	0	1	-7	4	1.00
3	-1	-3	1	1	-5	1.35
4	0	-4	0	6	0	1.80
5	1	-3	-1	1	5	2.60
6	2	0	-1	-7	-4	3.60
7	3	5	1	3	1	5.45
$\sum \xi_{ij}^2$	28	84	6	154	84	$16.55 = \sum y_j$
$\sum \xi_{ij}^2 y_j$	20.55	11.95	0.85	1.15	0.55	$56.0475 = \sum y_j^2$
λ	1	1	$\frac{1}{6}$	$\frac{7}{12}$	$\frac{7}{20}$	

Heraf fås

$$\begin{aligned}\sum (y_i - \bar{y})^2 &= 56.0475 - \frac{16.55^2}{7} \\ &= 56.0475 - 39.1289 \\ &= 16.9186\end{aligned}$$

$$\begin{aligned}\hat{\alpha} &= \tfrac{16.55}{7} = 2.36 \\ \hat{\beta}_1 &= \tfrac{20.55}{28} = 0.7339 \quad \text{SAK}_{1.\text{grad}} = \tfrac{20.55^2}{28} = 15.0822 \\ \hat{\beta}_2 &= \tfrac{11.95}{84} = 0.1423 \quad \text{SAK}_{2.\text{grad}} = \tfrac{11.95^2}{84} = 1.7000 \\ \hat{\beta}_3 &= \tfrac{0.85}{6} = 0.1417 \quad \text{SAK}_{3.\text{grad}} = \tfrac{0.85^2}{6} = 0.1204 \\ \hat{\beta}_4 &= \tfrac{1.15}{154} = 0.0075 \quad \text{SAK}_{4.\text{grad}} = \tfrac{1.15^2}{154} = 0.0086 \\ \hat{\beta}_5 &= \tfrac{0.55}{84} = 0.0065 \quad \text{SAK}_{5.\text{grad}} = \tfrac{0.55^2}{84} = 0.0036\end{aligned}$$

Vi samler resultaterne i nedenstående skema.

Vi ser, at leddene af 1., 2., og 3. grad er signifikante og de to følgende ikke signifikante, så vi bruger et polynomium af 3.die grad til beskrivelse.

Varition	SAK	f	S^2	Test	F-fraktion
Total	16.9186	6			
1. grad	15.0822	1	15.0822		
Residual 1	1.8364	5	0.3673	41.06	99.8%
2. grad	1.7000	1	1.7000		
Residual 2	0.1364	4	0.0341	49.85	99.7%
3. grad	0.1204	1	0.1204		
Residual 3	0.0160	3	0.0053	22.72	98.0%
4. grad	0.0086	1	0.0086		
Residual 4	0.0074	2	0.0037	2.32	75.0%
5. grad	0.0036	1	0.0036		
Residual 5	0.0038	1	0.0038	0.95	$< 50.0\%$

Af rekursionsformlen 4.3 og 4.3 fås - da $n = 7$ -

$$\begin{aligned}
\xi_1(t) &= t - 4 \\
\xi_2(t) &= (t-4)^2 - \frac{48}{12} \\
&= t^2 - 8t + 12 \\
\xi_3(t) &= (t-4)(t^2 - 8t + 12) - \frac{4 \cdot 45}{4 \cdot 15}(t-4) \\
&= t^3 - 12t^2 + 41t - 36.
\end{aligned}$$

Da $\lambda_1 = 1$, $\lambda_2 = 1$ og $\lambda_3 = 1/6$ får vi følgende estimerede polynomium

$$\begin{aligned}
\hat{\mu}(t) &= 2.36 + 1 \cdot \hat{\beta}_1 \xi_1(t) + 1 \cdot \hat{\beta}_2 \xi_2(t) + 1/6 \hat{\beta}_3 \xi_3(t) \\
&= 0.0236t^3 - 0.1409t^2 + 0.5631t + 0.2818.
\end{aligned}$$

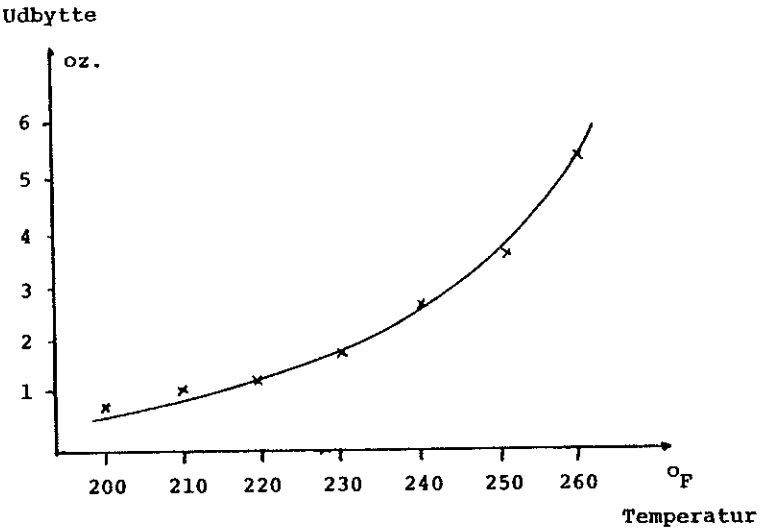
Da

$$t_i = \frac{\tau_i - 190}{10},$$

får vi et udtryk, hvori de oprindelige temperaturer indgår, ved blot at indsætte denne relation i udtrykket for $\hat{\mu}(t)$. Det giver

$$g(\tau) = 0.000024\tau^3 - 0.014861\tau^2 + 3.147610\tau_i - 223.15440.$$

Dette estimerede polynomium er sammen med de oprindelige data anført i nedenstående figur. ♦



Figur 4.4: Sammenhængen mellem temperatur og udbytte ved den i eksempel 4.2 angivne proces.

4.3 Valg af den "bedste" regressionsligning

I dette afsnit vil vi beskæftige os med problemet at vælge en passende (lille) mængde af uafhængige variable, som giver mulighed for en rimelig beskrivelse af vort datamateriale.

4.3.1 Problemstillingen

Hvis man e.g. er i den (kedelige) situation ikke at kunne formulere en på de fysiske forhold begrundet model for det fænomen, man studerer, vil man ofte som en sidste udvej ty til blot at registrere alle de størrelser, man overhovedet mener kan være af betydning for ens måleværdier. Hvis man så laver regression efter f.eks. polynomier i disse uafhængige variable (ud fra en Taylor-approximations-synsvinkel), vil man meget hurtigt komme op på et enormt antal led i ens regressionsligning. Har man 10 "grund"-variable $x_1, \ldots, x_{10}$, vil der i et almindeligt 2.dengrads polynomium i disse findes 66 led. Går vi op til 3.die grad får vi i størrelsesordenen 150 led. Udtryk der indeholder så mange led vil - hvis det overhovedet er muligt at estimere alle parametrene - være meget tunge at arbejde med. Hvis man e.g. ønsker at bestemme optimale produktionsbetingelser for en kemisk proces, kan man estimere responsfladen og finde maksimum for denne. Dette vil være en endog overordentlig vanskelig opgave, hvis der er mange variable involveret. Man vil derfor søge at **finde et væsentligt mindre antal led, der giver en tilpas "god" beskrivelse af variationen i materialet.** (Jævnfør også afsnittet om ridge

regression).

Det er dog vigtigt her at gøre sig klart, at et udtryk man kommer frem til ved de metoder, vi skal omtale, må benyttes med varsomhed. Der vil (formentlig) være tale om et udtryk, der beskriver de **foreliggende** data udmærket. Om metoden er velegnet til at **forudsige fremtidige** observationer, vil afhænge af, om udtrykket også afspejler de fysiske forhold tilstrækkeligt godt. En måde at få opklaret dette problem på, er i første omgang alene at basere estimationen på e.g. halvdelen af datamaterialet, og så sammenligne den resterende halvdel med den estimerede model. Hvis overensstemmelsen her er tilstrækkeligt stor, har man fået en indikation for, at modellen ikke er uanvendelig som forudsigelsesmodel.

Vi vil bruge et enkelt gennemgående eksempel som illustration af de metoder, vi vil omtale. For at det skal være muligt at overskue (og eventuelt kontrollere) de enkelte beregninger, har vi kun udtaget en meget lille del af det oprindelige materiale. Man skal derfor ikke vurdere metodernes egnethed v.h.a. eksemplet, men kun anvende det som illustration af principperne og gangen i disse. Data er nogle sammenhørende målinger af kvaliteten y af et fødeadditiv (målt ved viskositeten) og nogle produktionsparametre x_1, x_2 og x_3 (tryk, temperatur og neutraliseringsgrad). For at lette beregningerne er data kodede, i.e. der er substraheret nogle konstanter og divideret med passende tal. Vi har følgende målinger

y	x_1	x_2	x_3
4.9	0	0	2
3.0	1	0	1
0.2	1	1	0
2.9	1	2	2
6.4	2	1	2

Erfaringen har vist, at det inden for et passende lille variationsområde af produktionsparametrene er rimeligt at regne med, at kvaliteten afhænger lineært af disse. Vi anvender derfor modellen

$$\mathrm{E}(Y|\mathbf{x}) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

eller skrevet på matrixform

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 2 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 2 & 2 \\ 1 & 2 & 1 & 2 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{bmatrix},$$

$\varepsilon \in \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$.

Vi har i appendix (p. 192) anført samtlige 2^3 regressionsanalyser med y som afhængig variabel og en eller flere af x'erne som uafhængige variable. Der er følgende mulige modeller

$$
\begin{array}{lllllllll}
M & : & \mathrm{E}(Y) & = & \alpha & + & \beta_1 x_1 & + & \beta_2 x_2 & + & \beta_3 x_3 \\
H_{12} & : & \mathrm{E}(Y) & = & \alpha & + & \beta_1 x_1 & + & \beta_2 x_2 & & \\
H_{13} & : & \mathrm{E}(Y) & = & \alpha & + & \beta_1 x_1 & + & & & \beta_3 x_3 \\
H_{23} & : & \mathrm{E}(Y) & = & \alpha & + & & & \beta_2 x_2 & + & \beta_3 x_3 \\
H_1 & : & \mathrm{E}(Y) & = & \alpha & + & \beta_1 x_1 & & & & \\
H_2 & : & \mathrm{E}(Y) & = & \alpha & + & & & \beta_2 x_2 & & \\
H_3 & : & \mathrm{E}(Y) & = & \alpha & + & & & & & \beta_3 x_3 \\
H_0 & : & \mathrm{E}(Y) & = & \alpha & & & & & &
\end{array}
$$

Vi anfører for hver af disse 8 modeller **estimatorerne** for α og β'erne, vi finder projektionen af observationsvektoren på det til modellen svarende underrum, vi bestemmer residualvektoren, residualvektorens kvadrerede længde (residualkvadratafvigelsessummen), variansskøn og den multiple korrelationskoefficient. Dernæst anfører vi **variansanalyseskemaer** for de mulige sekvenser af successive testninger af hypoteser om, at middelværdivektoren tilhører stadigt mindre (lavere dimension underrum i kæder som

$$M \supseteq H_{12} \supseteq H_2 \supseteq H_0.$$

Ovenstående kæde af underrum svarer til successiv testning af hypoteserne

$$\beta_3 = 0, \quad \beta_1 = 0, \quad \beta_2 = 0.$$

Der er 6(=3!)mulige skemaer af denne art. Endelig anføres nogle **partielle korrelationsmatricer**. Sætter vi $y = x_4$ defineres dispersionmatricen (den empiriske) som bekendt ved, at det (i,j)'te element er

$$S_{ij} = \frac{1}{n-1} \sum_{\mu} (x_{i\mu} - \bar{x}_i)(x_{j\mu} - \bar{x}_j).$$

Korrelationsmatricens (i,j)'te element er da

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii} s_{jj}}}.$$

Ved hjælp af formlen p. 84 i afsnit 2 beregnes dernæst de partielle korrelationer for givet x_3 og for givet x_2, x_3.

Vi er nu rustede til at omtale nogle af de mest anvendte måder til at udvælge enkelte uafhængige variable til beskrivelse af variationen i den afhængige variable.

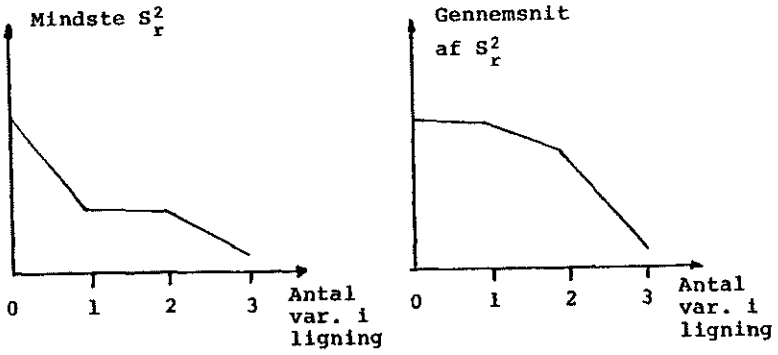
4.3.2 Undersøgelse af samtlige regressioner

Denne metode kan selvsagt kun anvendes, hvis der er rimeligt få variable.

Vi samler resultatet fra appendicet i følgende skema

Model	Multiple R^2	Residualv. S_r^2	Gennemsn. af S_r^2
H_0 : $E(Y) = \alpha$	0	5.47	5.47
H_1 : $E(Y) = \alpha + \beta_1 x_1$	5.1%	6.91	5.35
H_2 : $E(Y) = \alpha + \beta_2 x_2$	3.8%	7.01	
H_3 : $E(Y) = \alpha + \beta_3 x_3$	70.8%	2.13	
H_{12} : $E(Y) = \alpha + \beta_1 x_1 + \beta_2 x_2$	15.3%	9.26	4.68
H_{13} : $E(Y) = \alpha + \beta_1 x_1 + \beta_3 x_3$	76.0%	2.63	
H_{23} : $E(Y) = \alpha + \beta_2 x_2 + \beta_3 x_3$	80.4%	2.14	
M : $E(Y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$	97.1%	0.634	0.634

Et blik på de multiple korrelationskoefficienter viser, at der ikke er vundet så afgørende meget ved at gå fra 1 variabel (x_3) op til 2 variable. Det afgørende spring sker først ved at gå op til fulde 3 variable. Overvejelser i disse baner fører til, at man vil foretrække alene at anvende x_3 d.v.s. modellen $\mathrm{E}(Y|x) = \alpha + \beta_3 x_3$. Denne afgørelse bliver yderligere bestyrket ved at se på residualvariansen S_r^2. Vi ser da, at S_r^2 for den "bedste" ligning i en variabel er mindre end for den "bedste" ligning i 2 variable, hvilket kraftigt indicerer, at vi bør nøjes med en variabel (eller tage alle tre). Ser vi foruden på de mindste S_r^2 også på de gennemsnitlige værdier og afbilder dem efter antallet af indgående variable, får vi grafer som



Dette antyder også, at antallet af variable i en ligning bør være enten 1 eller 3 (der opnås ingen væsentlige forbedringer ved at gå fra 1 til 2).

Ser man alene på grafen med de gennemsnitlige værdier er det ikke indlysende, om man overhovedet bør inddrage nogen uafhængig variabel. Man kan derfor teste om β_3

i modellen H_3 ($\mathrm{E}(y|x) = \alpha + \beta_3 x_3$) kan antages at være 0. Teststørrelsen er

$$\frac{||p_{H_0}(\mathbf{y}) - p_{H_3}(\mathbf{y})||^2/1}{||\mathbf{y} - p_{H_3}||^2/3} = \frac{21.868 - 6.38}{6.38/3} \simeq 7.28.$$

Vi vil derfor få forkastet $\beta_3 = 0$ på alle niveauer større end 8%.

Som en sammenfatning af disse (løse) betragtninger vil vi konkludere, at vi anvender modellen H_3:

$$\mathrm{E}(Y|x) = \alpha + \beta_3 x_3 \simeq 0.4 + 2.2 x_3.$$

(Her betegner $\simeq$ estimeret til). Skønnet over usikkerheden (variansen) på målingerne er (estimeret med 3 frihedsgrader)

$$s^2 = 2.13.$$

BEMÆRKNING 4.4. Det må her tilføjes, at ideen med at se på gennemsnittene af residualvarianserne forekommer noget tvivlsom. Den er medtaget fordi metoden nyder en vis udbredelse - i det mindste i litteraturen. ▼

4.3.3 Backwards elimination

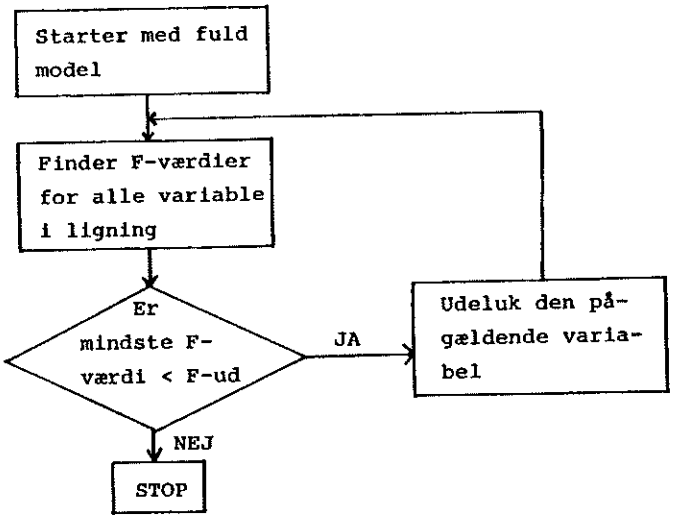
Denne metode er langt mere økonomisk, hvad angår regnetid, end foregående. Man starter her med den fuldstændige model og undersøger så hvilken af koefficienterne, der har den mindste F-værdi for et test af hypotesen, at koefficienten er 0.

Denne variable udelukkes, og proceduren gentages så med de resterende $k - 1$ variable etc.

Man kan så stoppe proceduren, når ingen af de tilbageværende variable har en F-værdi, der er mindre end $1 - \alpha$ fraktilen i den relevante F-fordeling.

Vi kan anskueliggøre fremgangsmåden ved hjælp af vor eksempel. Vi samler data i omstående skema.

Af skemaet fremgår, at vi også her vil ende med modellen H_3: $\mathrm{E}(y) = \alpha + \beta_3 x_3$, når vi arbejder med et α, der er større end 8%.

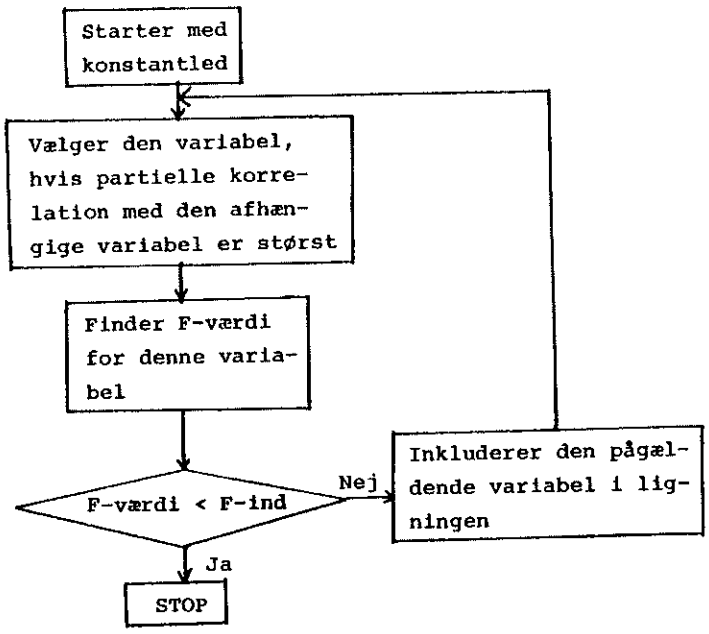


Figur 4.5: Flow diagram for Backwards-elimination procedure i trinvis regressionsanalyse.

Trin	F-værdi for test af $\beta_i = 0$	Fraktil i F-fordel.
Model : $E(Y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$		
1	$\beta_1 : \frac{3.625/1}{0.634/1} = 5.76$	$= F(1,1)_{0.71}$
	$\beta_2 : \frac{4.621/1}{0.634/1} = 7.29$	$= F(1,1)_{0.72}$
	$\beta_3 : \frac{17.879/1}{0.634/1} = 28.20$	$= F(1,1)_{0.86}$
Udelader x_1 : model : $E(Y) = \alpha + \beta_2 x_2 + \beta_3 x_3$		
2	$\beta_2 : \frac{2.095/1}{4.285/2} = 0.98$	$= F(1,2)_{0.55}$
	$\beta_3 : \frac{16.757/1}{4.285/2} = 7.82$	$= F(1,2)_{0.88}$
Udelader x_2 : model : $E(Y) = \alpha + \beta_3 x_3$		
3	$\beta_3 : \frac{15.488/1}{6.38/3} = 7.28$	$= F(1,3)_{0.92}$

Ulempen ved denne metode er, at vi skal løse den fuldstændige regressionsmodel, hvilket kan være besværligt, hvis der er mange uafhængige variable.

Dette problem bliver der taget højde for ved den næste procedure.



Figur 4.6: Flowdiagram for Forward-selection procedure i trinvis regressionsanalyse.

4.3.4 Forward selection

Ved denne procedure starter man med kun at have konstantleddet i ligningen. Dernæst vælger man den uafhængige variable, der er mest korreleret med den afhængige variable. Man udfører et F-test for, om dennes koefficient er signifikant $\neq 0$. Hvis ja, drages den med ind i modellen.

Blandt de resterende uafhængige variable udvælges nu den, der har den største (numerisk) partielle korrelationskoefficient med den afhængige variable givet de (n) variable, der allerede er i ligningen. Man udfører et F-test for om denne nye variable har bidraget til reduktion af residualvariansen, i.e. om koefficienten til den er $\neq 0$. Hvis ja, fortsættes som før, hvis nej, standses analysen.

I vort eksempel bliver gangen

1) Af korrelationsmatricen (p. 196) ses, at x_3 har den højeste korrelationskoefficient med y, nemlig 0.8416. Vi tester, om β_3 i modellen $\mathrm{E}(Y) = \alpha + \beta_3 x_3$ kan antages at være 0. Vi har teststørrelsen (se p. 196).

$$\frac{15.488/1}{6.38/3} = 7.28 \simeq \mathrm{F}(1,3)_{0.92}.$$

Hvis vi regner med $\alpha = 10\%$ fortsættes (da vi da forkaster $\beta_3 = 0$).

2) Af den partielle korrelationsmatrix givet x_3 (p. 197) ses, at den variable, der, givet x_3 er med i ligningen, har den største partielle korrelationskoefficient med y'erne er x_2 ($\rho_{x_2y|x_3} = -0.5728$). Vi inddrager x_2 og undersøger, om β_2 i modellen

$$\mathrm{E}(y) = \alpha + \beta_2 x_2 + \beta_3 x_3$$

kan antages at være 0. Vi har teststørrelsen (se p. 196)

$$\frac{2.095/1}{4.2855/2} = 0.98 \simeq \mathrm{F}(1,2)_{0.55}.$$

Da vi regnede med $\alpha = 10\%$, er denne størrelse altså ikke signifikant forskellig fra 0, og vi slutter analysen her og uden at medtage x_2. Den resulterende model er

$$\mathrm{E}(Y) = \alpha + \beta_3 x_3,$$

hvor α og β estimeres som tidligere. Vi bemærker her specielt, at x_1 slet ikke har været med i ligningen.

BEMÆRKNING 4.5. Hvis vi havde opereret med $\alpha = 50\%$, ville vi have fortsat analysen og betragtet de partielle korrelationer givet x_2 og x_3. Ifølge matricen p. 197 er den partielle korrelationskoefficient mellem y og x_1 givet x_2 og x_3 er i ligningen

$$\rho_{x_1y|x_2x_3} = 0.8956.$$

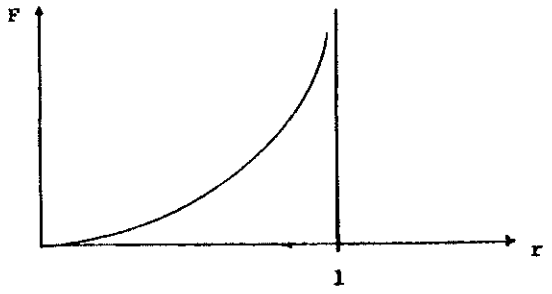
Nu er x_1 den eneste tiloversblevne variable, så derfor er den trivielt den af de tiloversblevne, der har de største partielle korrelationer med y. Vi drager x_1 med ind i ligningen og undersøger, om β_1 i modellen $\mathrm{E}(y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ er signifikant forskellig fra 0. Teststørrelsen er (p. 196)

$$\frac{3.652/1}{0.634/1} = 5.76 \simeq \mathrm{F}(1,1)_{0.71}.$$

I dette tilfælde har vi set, at ligningen blev udvidet væsentligt ved at ændre α. Det er vigtigt at gøre sig klar, at ændringer i α kan have drastiske ændringer i den resulterende model til følge. ▼

BEMÆRKNING 4.6. Det i hvert trin at vælge den variable, der har den største partielle korrelation med den afhængige variable, er ensbetydende med at vælge de variable, der har den største F-værdi i de partielle F-tests. Dette resultat følger af, at sammenhængen mellem den partielle korrelationskoefficient og F-teststørrelsen er af formen

$$\mathrm{F} = \mathrm{g}(r) = \frac{r^2}{1-r^2} \cdot f,$$



hvor f er frihedsgradsantallet for nævneren (jvf. p. 162). Denne sammenhæng er åbenbart voksende

Hvis vi e.g. i trin 2 vil beregne F-teststørrelsen ud fra korrelationsmatricen, får vi

$$\mathrm{F} = \frac{(-0.5728)^2}{1-(-0.5728)^2} \cdot 2 = 0.98.$$

Dette ses endvidere, at omtalte kriterium er ensbetydende med i hvert enkelt trin hele tiden at indtage den variable, der bevirker den største reduktion i residualkvadratafvigelsessummen. ▼

BEMÆRKNING 4.7. I mange af de eksisterende standardregressionsprogrammer er det ikke muligt at specificere en α-værdi. Man må i stedet angive et fast tal som grænse for de F-teststørrelser, man vil acceptere respektive forkaste. Man må da ved at se i en tabel over F-fraktiler finde en passende værdi. Ønsker man e.g. $\alpha \simeq 5\%$, ser man, at man bør vælge værdien 4, idet

$$\mathrm{F}(1,n)_{0.95} \simeq 4,$$

for rimeligt store værdier af n. ▼

"Forward selection" metoden udmærker sig frem for "Backward elimination" metoden ved, at vi ikke behøver at regne den totale ligning ud. Den største ulempe ved metoden er nok, at der ikke tages hensyn til, at nogle variable kan være blevet overflødiggjort ved at andre senere er kommet ind. Hvis vi f.eks. tænker os, at $x_1 = ax_2 + bx_3$ (ca.), og at x_1 er valgt som den mest betydende variable. Hvis vi så senere i analysen også inddrager x_2 og x_3, er det klart, at vi ikke længere har "brug" for x_1. Den bør derfor udelukkes. Det sker ved den sidste metode, vi omtaler:

4.3.5 Stepwise regression

Navnet er dårligt, idet man med lige så fuld ret ville kunne benævne de to foregående metoder med dette navn. Der er også mange forfattere, der bruger navnet stepwise regression som en fællesbetegnelse for en række forskellige procedurer. I nærværende fremstilling tænker vi helt specifikt på følgende metode.

Udvælgelsen af den variable, der skal ind i ligningen foregår som ved forward selection proceduren; men i hvert enkelt trin undersøger man så hver af de variable, der er i ligningen, som om de var de sidst tilføjede variable. Man beregner så på denne måde en F-teststørrelse for alle variable,der er i ligningen. Hvis nogle af disse er mindre end $1-\alpha$-fraktilen i den relevante F-fordeling, udelukkes de pågældende variable.

Hvis vi ser på vort standardeksempel, får vi følgende trin ($\alpha_{\text{ind}} = 50\%, \alpha_{\text{ud}} = 40\%$).

1) x_3 tages ind som ved forward selection proceduren, og vi tester om β_3 er signifikant forskellig fra 0. Teststørrelsen og konklusion bliver som før.

2) Vi tager nu x_2 ind. Vi danner det partielle F-test for β_2 (i modellen $\mathrm{E}(Y) = \alpha + \beta_2 x_2 + \beta_3 x_3$):

$$x_2: \quad \text{F-værdi} = \frac{2.095/1}{4.285/2} = 0.98 \simeq \mathrm{F}(1,2)_{0.55}.$$

Dernæst laver vi et partielt F-test for β_3 (i modellen $\mathrm{E}(Y) = \alpha + \beta_2 x_2 + \beta_3 x_3$). Ved hjælp af skemaet p. 196 fås, at

$$x_3: \quad \text{F-værdi} = \frac{16.757/1}{4.285/2} = 7.82 \simeq \mathrm{F}(1,2)_{0.88}.$$

3) Vi fjerner igen x_2 fra ligningen da $0.55 < 0.60$. Forskellen i dette trin mellem forward selection proceduren og stepwise proceduren er, at vi også beregner en F-værdi for x_3 og dermed åbner mulighed for at x_3 igen elimineres fra ligningen. Dette var ikke muligt ved den rene forward selection procedure.

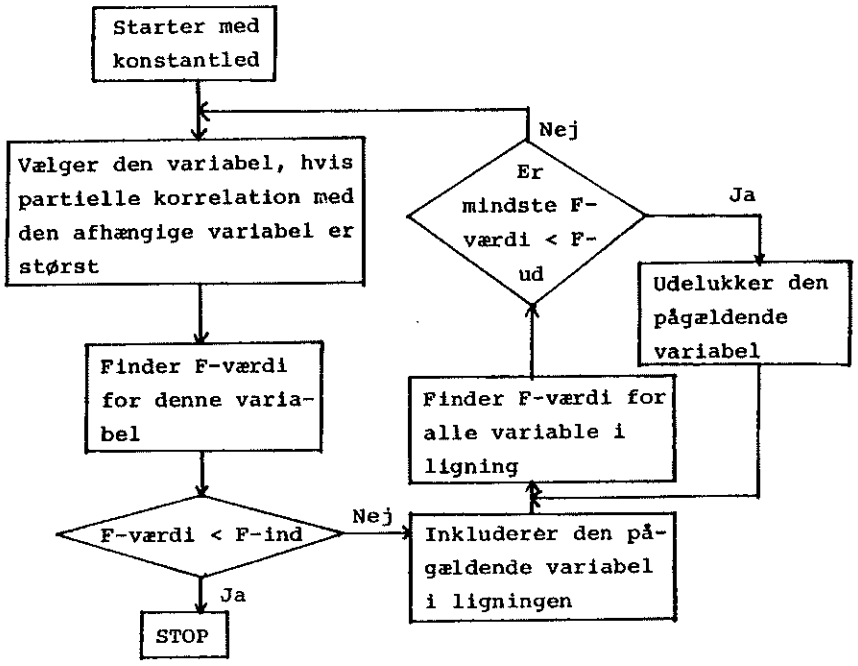
4) Den eneste tilbageværende variable er x_1. Den har en partiel F-værdi på

$$x_1: \quad \text{F-værdi} = \frac{1.125/1}{5.255/2} = 0.43 < \mathrm{F}(1,2)_{0.50},$$

den kommer overhovedet ikke ind i ligningen.

Analysen stopper, og vi har modellen

$$\mathrm{E}(Y) = \alpha + \beta_3 x_3.$$



Figur 4.7: Flowdiagram for Stepwise-Regression procedure i trinvis regressionsanalyse.

BEMÆRKNING 4.8. Grunden til, at vi undersøgte den partielle F-værdi under 2, men ikke under 4 er, at x_1 overhovedet ikke kommer ind i ligningen, da

$$0.43 < \mathrm{F}(1,2)_{0.50} = \mathrm{F}_{1-\alpha_{ind}}.$$

Derimod kom x_2 ind i ligningen, da

$$0.98 < \mathrm{F}(1,2)_{0.55} > \mathrm{F}_{1-\alpha_{ind}}.$$

▼

BEMÆRKNING 4.9. Som under afsnittet om forward selection proceduren kan vi konstatere, at man ofte bliver tvunget til at arbejde med faste F-værdier i stedet for $1-\alpha$ fraktiler. Når man ikke anvender samme niveau ved afgørelsen af, om man vil tage flere variable ind, som man bruger ved undersøgelse af, om nogle variable skal ud, vil man ofte gøre den sidste ca. halvt så stor som den første, i.e.

$$\text{F-ud af ligning} = 1/2\text{F-ind i ligning}.$$

(Dette er det modsatte af det, vi har anvendt i eksemplet). ▼

4.3.6 Nogle eksisterende programmer

Vi skal ingenlunde give en oversigt over, hvad der findes af regressionsanalyseprogrammer, men blot nævne 3, der alle er let tilgængelige på NEUCC.

Der er for det første **SSP-programmet** STEPR, som kalder en del subroutiner fra SSP biblioteket. Det er et rent **forward selection program**. Det er ikke umiddelbart muligt at tvinge en regressionsflade gennem 0. Der er mulighed for at få tabel over residualer. De af STEPR kaldte routiner fra SSP-biblioteket er lagt ind under WATFIV-systemet.

BMDO2R er et "stepwise" program. Det er med dette system hel trivielt at tvinge en regressionsflade gennem 0. Det er endvidere muligt at få såvel tabeller over residualer som plots af residualer mod input-variable. Endelig kan man få udskrevet tabel over analysens forskellige trin.

Det af **H. Spliid** udarbejdede program **REGR**, er et "stepwise" program, der er kombineret med en **backwards elimination** procedure. Det vil derfor være normalt her at anvende mindre F-værdier for ind- og udtagning af variable under den trinvise regression. Man får da en ligning indeholdende temmeligt mange variable, og på denne kan man anvende en backwards eliminatin procedure. Denne metode er nok en af de mere velegnede til at finde gode "prediktorer" blandt et stort antal variable. I de nyeste version af REGR er det dels muligt at tvinge regressionsfladen gennem 0, og dels er det

muligt at udføre en simpel vægtet regression. Der findes særlige residualplotrutiner til programmet, og endvidere findes der en version af REGR under WATFIV systemet.

En meget omfattende samling regressionsrutiner findes i **SAS-systemet**. Foruden Forward og Hackwards proceduren findes en stepwise procedure, en såkaldt "maximum R^2-improvement" procedure, en procedure der udfører samtlige regressioner m.fl.

For alle de nævnte programmer gælder, at man på forhånd har mulighed for at angive visse variable, som man altid ønsker medtaget i regressionsligningen.

4.3.7 Numerisk appendix

Vi anfører i dette appendix en udregning af de størrelser, der er anvendt i de tidligere afsnit. Det skulle ikke være nødvendigt at gennemgå alle disse udregninger, men de er anført, for at man ved hjælp af disse skal kunne checke sin forståelse af de forskellige principper.

A. DATA:

y	x_1	x_2	x_3
4.9	0	0	2
3.0	1	0	1
0.2	1	1	0
2.9	1	2	2
6.4	2	1	2

B. Grundmodel: $\mathrm{E}(Y) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$ eller

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 2 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 2 & 2 \\ 1 & 2 & 1 & 2 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \end{bmatrix}$$

$$\varepsilon \in \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I})$$

C. Estimatorer i submodeller

i) **Model M:** $\mathrm{E}(\boldsymbol{Y}) = \boldsymbol{\alpha} + \boldsymbol{\beta_1 X_1} + \boldsymbol{\beta_2 X_2} + \boldsymbol{\beta_3 x_3}$

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} -0.175 \\ 1.450 \\ -1.400 \\ 2.375 \end{bmatrix}; p_M(\mathbf{y}) = \begin{bmatrix} 4.575 \\ 3.650 \\ -0.125 \\ 3.225 \\ 6.075 \end{bmatrix}; \mathbf{y} - p_M(\mathbf{y}) = \begin{bmatrix} 0.325 \\ -0.650 \\ 0.325 \\ -0.325 \\ 0.325 \end{bmatrix}$$

$$\frac{1}{5-4}\|\mathbf{y}-p_M(\mathbf{y})\|^2 = \frac{0.845}{1} = 0.845$$
$$R^2 = \frac{21.868 - 0.633750}{21.868} = 97.1\%$$

ii) **Model H_{12} : $\mathbf{E}(Y) = \alpha + \beta_1 x_1 + \beta_2 x_2$**

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 3.026 \\ 1.243 \\ -0.987 \end{bmatrix}; p_{H_{12}}(\mathbf{y}) = \begin{bmatrix} 3.026 \\ 4.269 \\ 3.282 \\ 2.295 \\ 4.525 \end{bmatrix}; \mathbf{y} - p_{H_{12}}(\mathbf{y}) = \begin{bmatrix} 1.874 \\ -1.269 \\ -3.082 \\ 0.605 \\ -1.875 \end{bmatrix}$$

$$\frac{1}{5-3}\|\mathbf{y}-p_{H_{12}}(\mathbf{y})\|^2 = \frac{18.512611}{2} = 9.2563$$
$$R^2 = \frac{21.868 - 18.512611}{21.868} = 15.3\%$$

iii) **Model H_{13}: $\mathbf{E}(Y) = \alpha + \beta_1 x_1 + \beta_3 x_3$**

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} -0.350 \\ 0.750 \\ 2.200 \end{bmatrix}; p_{H_{13}}(\mathbf{y}) = \begin{bmatrix} 4.05 \\ 2.60 \\ 0.40 \\ 4.80 \\ 5.55 \end{bmatrix}; \mathbf{y} - p_{H_{13}}(\mathbf{y}) = \begin{bmatrix} 0.85 \\ 0.40 \\ -1.20 \\ -1.90 \\ 0.85 \end{bmatrix}$$

$$\frac{1}{5-3}\|\mathbf{y}-p_{H_{13}}(\mathbf{y})\|^2 = \frac{5.2250}{2} = 2.6275$$
$$R^2 = \frac{21.868 - 5.2550}{21.868} = 76.0\%$$

iv) **Model H_{23}: $\mathbf{E}(Y) = \alpha + \beta_2 x_2 + \beta_3 x_3$**

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} 0.945 \\ -0.872 \\ 2.309 \end{bmatrix}; p_{H_{23}}(\mathbf{y}) = \begin{bmatrix} 5.563 \\ 3.254 \\ 0.073 \\ 3.819 \\ 4.691 \end{bmatrix}; \mathbf{y} - p_{H_{23}}(\mathbf{y}) = \begin{bmatrix} -0.663 \\ -0.254 \\ 0.127 \\ -0.919 \\ 1.709 \end{bmatrix}$$

$$\frac{1}{5-3}\|\mathbf{y}-p_{H_{23}}(\mathbf{y})\|^2 = \frac{4.285456}{2} = 2.1427$$
$$R^2 = \frac{21.868-4.2855}{21.868} = 80.4\%$$

v) **Model H_1 : $\mathbf{E}(Y) = \alpha + \beta_1 x_1$**

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta}_1 \end{bmatrix} = \begin{bmatrix} 2.73 \\ 0.75 \end{bmatrix}; p_{H_1}(\mathbf{y}) = \begin{bmatrix} 2.73 \\ 3.48 \\ 3.48 \\ 3.48 \\ 4.23 \end{bmatrix}; \mathbf{y} - p_{H_1}(\mathbf{y}) = \begin{bmatrix} 2.17 \\ -0.48 \\ -3.28 \\ -0.58 \\ 2.17 \end{bmatrix}$$

$$\frac{1}{5-2}\|\mathbf{y}-p_{H_1}(\mathbf{y})\|^2 = \frac{20.7430}{3} = 6.9143$$
$$R^2 = \frac{21.868-20.743}{21.868} = 5.1\%$$

vi) **Model H_2: $\mathbf{E}(Y) = \alpha + \beta_2 x_2$**

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 3.914 \\ -0.543 \end{bmatrix}; p_{H_2}(\mathbf{y}) = \begin{bmatrix} 3.914 \\ 3.914 \\ 3.371 \\ 2.828 \\ 3.371 \end{bmatrix}; \mathbf{y} - p_{H_2}(\mathbf{y}) = \begin{bmatrix} 0.986 \\ -0.914 \\ -3.171 \\ 0.072 \\ 3.029 \end{bmatrix}$$

$$\frac{1}{5-2}\|\mathbf{y}-p_{H_2}(\mathbf{y})\|^2 = \frac{21.042858}{3} = 7.0143$$
$$R^2 = \frac{21.868-21.043}{21.868} = 3.8\%$$

vii) **Model H_3: $\mathbf{E}(Y) = \alpha + \beta_3 x_3$**

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} 0.4 \\ 2.2 \end{bmatrix}; p_{H_3}(\mathbf{y}) = \begin{bmatrix} 4.8 \\ 2.6 \\ 0.4 \\ 4.8 \\ 4.8 \end{bmatrix}; \mathbf{y} - p_{H_3}(\mathbf{y}) = \begin{bmatrix} 0.1 \\ 0.4 \\ -0.2 \\ -1.9 \\ 1.6 \end{bmatrix}$$

$$\frac{1}{5-2}\|\mathbf{y}-p_{H_3}(\mathbf{y})\|^2 = \frac{6.38}{3} = 2.1267$$
$$R^2 = \frac{21.868-6.38}{21.868} = 70.8\%$$

viii) **Model H_0: $\mathrm{E}(Y) = \alpha$**

$$\hat{\alpha} = 3.48$$

$$p_{H_0}(\mathbf{y}) = \begin{bmatrix} 3.48 \\ 3.48 \\ 3.48 \\ 3.48 \\ 3.48 \end{bmatrix}; \mathbf{y} - p_{H_0}(\mathbf{y}) = \begin{bmatrix} 1.42 \\ -0.48 \\ -3.28 \\ -0.58 \\ 2.92 \end{bmatrix}$$

$$\frac{1}{5-1}\|\mathbf{y}-p_{H_0}(\mathbf{y})\|^2 = \frac{21.8680}{4} = 5.4670$$

D. Succesive testninger

1) $H \supseteq H_{12} \supseteq H_1 \supseteq H_0$ d.v.s. : $\beta_3 = 0, \beta_2 = 0, \beta_1 = 0$

Variation		SAK	fr.gr.
$H_0 - H_1$	$(\beta_1 = 0)$	$21.868 - 20.7430 = 1.125$	1
$H_1 - H_{12}$	$(\beta_2 = 0)$	$20.7430 - 18.5126 = 2.230$	1
$H - H_{12}$	$(\beta_3 = 0)$	$18.5126 - 0.6338 = 17.879$	1
M – obs		$0.6338 = 0.634$	1
H_0 – obs		21.868	4

2) $M \supseteq H_{12} \supseteq H_2 \supseteq H_0$ d.v.s. : $\beta_3 = 0, \beta_1 = 0, \beta_2 = 0$

Variation		SAK	fr.gr.
$H_0 - H_2$	$(\beta_2 = 0)$	$21.8680 - 21.0429 = 0.825$	1
$H_2 - H_{12}$	$(\beta_1 = 0)$	$21.0429 - 18.5126 = 2.530$	1
$H_{12} - M$	$(\beta_3 = 0)$	$18.5126 - 0.6338 = 17.879$	1
M – obs		$0.6338 = 0.634$	1
H_0 – obs		21.868	4

3) $M \supset H_{13} \supset H_1 \supset H_0$ d.v.s. : $\beta_2 = 0, \beta_3 = 0, \beta_1 = 0$

Variation		SAK	fr.gr.
$H_0 - H_1$	$(\beta_1 = 0)$	$21.8680 - 20.7430 = 1.125$	1
$H_1 - H_{13}$	$(\beta_3 = 0)$	$20.7430 - 5.2550 = 15.488$	1
$H_{13} - M$	$(\beta_2 = 0)$	$5.2550 - 0.6338 = 4.621$	1
M – obs		$0.6338 = 0.634$	1
H_0 – obs		21.868	4

4) $M \supseteq H_{13} \supseteq H_3 \supseteq H_0$ d.v.s. : $\beta_2 = 0, \beta_1 = 0, \beta_3 = 0$

Variation		SAK	fr.gr.
$H_0 - H_3$	$(\beta_3 = 0)$	$21.8680 - 6.38 = 15.488$	1
$H_3 - H_{13}$	$(\beta_1 = 0)$	$6.38 - 5.2550 = 1.125$	1
$H_{13} - M$	$(\beta_2 = 0)$	$5.2550 - 0.6338 = 4.621$	1
$M -$ obs		$0.6338 = 0.634$	1
$H_0 -$ obs		21.868	4

5) $M \supseteq H_{23} \supseteq H_2 \supseteq H_0$ d.v.s. : $\beta_1 = 0, \beta_3 = 0, \beta_2 = 0$

Variation		SAK	fr.gr.
$H_0 - H_2$	$(\beta_2 = 0)$	$21.8680 - 21.0429 = 0.825$	1
$H_2 - H_{23}$	$(\beta_3 = 0)$	$21.0429 - 4.2855 = 16.757$	1
$H_{23} - M$	$(\beta_1 = 0)$	$4.2855 - 0.6338 = 3.652$	1
$M -$ obs		$0.6338 = 0.634$	1
$H_0 -$ obs		21.868	4

6) $M \supset H_{23} \supset H_3 \supset H_0$ d.v.s. : $\beta_1 = 0, \beta_2 = 0, \beta_3 = 0$

Variation		SAK	fr.gr.
$H_0 - H_3$	$(\beta_3 = 0)$	$21.8680 - 6.38 = 15.488$	1
$H_3 - H_{23}$	$(\beta_2 = 0)$	$6.38 - 4.2855 = 2.095$	1
$H_{23} - M$	$(\beta_1 = 0)$	$4.2855 - 0.6338 = 3.652$	1
$M -$ obs		$0.6338 = 0.634$	1
$H_0 -$ obs		21.868	4

E. Dispersions- og korrelationsmatrix for data

$$\text{Dispersionsmatrix} = \frac{1}{5-1}\begin{pmatrix} 2 & 1 & 0 & 1.50 \\ 1 & 2.8 & 0.4 & -1.52 \\ 0 & 0.4 & 3.2 & 7.04 \\ 1.50 & -1.52 & 7.04 & 21.868 \\ x_1 & x_2 & x_3 & y \end{pmatrix} \begin{matrix} x_1 \\ x_2 \\ x_3 \\ y \end{matrix}$$

$$\text{korrelationsmatrix} = \begin{pmatrix} 1 & 0.4225 & 0 & 0.2268 \\ 0.4225 & 1 & 0.13393 & -0.1942 \\ 0 & 0.1336 & 1 & 0.8416 \\ 0.2268 & -0.1942 & 0.8416 & 1 \\ x_1 & x_2 & x_3 & y \end{pmatrix} \begin{matrix} x_1 \\ x_2 \\ x_3 \\ y \end{matrix}$$

F. Partielle korrelationer givet x_3:

$$\begin{pmatrix} 1 & 0.4225 & 0.2268 \\ 0.4225 & 1 & -0.1942 \\ 0.2268 & -0.1942 & 1 \end{pmatrix} - \begin{pmatrix} 0 \\ 0.1336 \\ 0.8416 \end{pmatrix} [1]^{-1} [\ 0 \quad 0.1336 \quad 0.8416\ ]$$
$$= \begin{pmatrix} 1 & 0.4225 & 0.2268 \\ 0.4225 & 0.9822 & -0.3066 \\ 0.2268 & -0.3066 & 0.2917 \end{pmatrix},$$

d.v.s. korrelationsmatricen er

$$\begin{pmatrix} 1 & 0.4263 & 0.4199 \\ 0.4263 & 1 & -0.5728 \\ 0.4199 & -0.5728 & 1 \\ x_1 & x_2 & y \end{pmatrix} \begin{matrix} x_1 \\ x_2 \\ y \end{matrix}$$

G. Partielle korrelationer givet x_2, x_3:

Først beregnet ud fra ovenstående partielle korralationsmatrix:

$$\begin{pmatrix} 1 & 0.4199 \\ 0.4199 & 1 \end{pmatrix} - \begin{pmatrix} 0.4263 \\ 0.5728 \end{pmatrix} [1]^{-1} [0.4263 - 0.5728] =$$
$$\begin{pmatrix} 0.8183 & 0.6641 \\ 0.6641 & 0.6718 \end{pmatrix},$$

hvorfor korrelationsmatricen bliver

$$\begin{pmatrix} 1 & 0.8956 \\ 0.8956 & 1 \end{pmatrix} \begin{matrix} x_1 \\ y \end{matrix}$$

Som kontrol beregnes den ud fra den oprindelige kovariansmatrix:

$$\begin{pmatrix} 2 & 1.50 \\ 1.50 & 21.868 \end{pmatrix} - \begin{pmatrix} 1 & 0 \\ -1.52 & 7.04 \end{pmatrix} \begin{pmatrix} 2.8 & 0.4 \\ 0.4 & 3.2 \end{pmatrix}^{-1} \begin{pmatrix} 1 & -1.52 \\ 0 & 7.04 \end{pmatrix}$$
$$= \begin{pmatrix} 2 & 1.50 \\ 1.50 & 21.868 \end{pmatrix} - \begin{pmatrix} 1 & 0 \\ -1.52 & 7.04 \end{pmatrix} \begin{pmatrix} 0.3636 & -0.0455 \\ -0.0455 & 0.3182 \end{pmatrix} \begin{pmatrix} 1 & -1.52 \\ 0 & 7.04 \end{pmatrix}$$
$$= \begin{pmatrix} 1.6363 & 2.3727 \\ 2.3727 & 4.2855 \end{pmatrix},$$

hvorfor den partielle korrelationsmatrix er

$$\begin{pmatrix} 1 & 0.8960 \\ 0.8960 & 1 \end{pmatrix} \begin{matrix} x_1 \\ y \end{matrix}$$

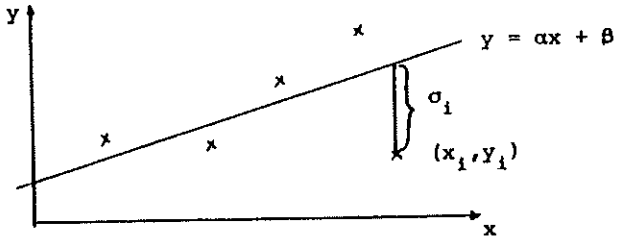
Afvigelserne i elementerne uden for diagonalen skyldes afrundingsfejl.

4.4 Andre regressionsmodeller og -løsninger

I dette afsnit skal vi dels betragte nogle mere "curve-fitting" orienterede problemer, dels nogle ulineære regressionsmodeller og endelig et alternativ til mindste kvadraters metode, nemlig den såkaldte ridge-regression.

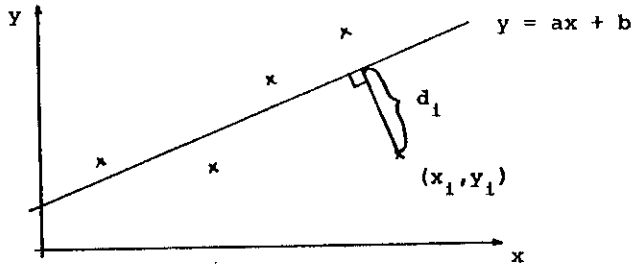
4.4.1 Ortogonal regression (lineær funktionel relation)

Ved den almindelige mindste kvadraters estimation af den regressionsflade minimaliserer man kvadratsummen af de lodrette afstande mellem regressionsfladen og de observerede punkter.



Almindelig regression: α og β bestemmes ved at minimalisere $\sum \sigma_i^2$.

Ofte vil man dog være i den situation, at det vil være mere rimeligt at minimalisere de ortogonale afstande, og så taler vi om **ortogonal regression** (ikke at forveksle med regression efter ortogonale polynomier).



Ortogonal regression: a og b bestemmes ved at minimalisere $\sum d_i^2$.

Lad der f.eks. foreligge variable $\mu_1, \ldots, \mu_p$, som tilfredsstiller en lineær relation

$$\alpha_0 + \alpha_1 \mu_1 + \cdots + \alpha_p \mu_p = 0, \tag{4.6}$$

d.v.s. de variable ligger i en hyperplan med ovenstående ligning. Vi er interesserede i at bestemme denne plan, i.e. i at bestemme $\alpha_0, \ldots, \alpha_p$. Antag, at det ikke er muligt at observere værdierne $\mu_1, \ldots, \mu_p$, men kun størrelser

$$X_{ji} = \mu_{ji} + Z_{ji}, \qquad j = 1, \ldots, p, \quad i = 1, \ldots, n,$$

hvor Z_{ji}'erne er stokastiske variable med middelværdi 0, og hvor $\mu_{1i}, \ldots, \mu_{pi}$, $i = 1, \ldots, n$, tilfredsstiller 4.6.

Estimationen af parametrene α_i på basis af et sådant sæt observationer kaldes i litteraturen ofte estimation af en **lineær funktionel relation**.

Det vil her være intuitivt rimeligt netop at anvende den hyperplan, der fås ved at minimalisere de ortogonale afstande ned til denne. Hvis Z_{ji}'erne er normalt fordelte med samme varians, kan det vises (se f.eks. [20] p. 392), at denne plan giver maksimum likelihood skønnene over α'erne.

Vi formulerer løsningen til problemet i

SÆTNING 4.3. Lad der være givet n punkter $\mathbf{x}_1, \ldots, \mathbf{x}_n \in R^p$. Koefficienterne i den hyperplan

$$\alpha_0 + \alpha_1 x_1 + \cdots + \alpha_p x_p = 0,$$

som minimaliserer kvadratsummen af de vinkelrette afstande fra punkterne til planen, er for $\alpha_1, \ldots, \alpha_p$'s vedkommende koordinaterne til en normeret egenvektor for den empiriske dispersionsmatrix for $\mathbf{x}$'erne svarende til den mindste egenværdi. Den sidste koefficient er givet ved

$$\alpha_0 = -\alpha_1 \bar{x}_1 - \cdots - \alpha_p \bar{x}_p.$$

▲

BEVIS 4.4. Vi forudsætter altså, at vi har observationerne

$$\begin{pmatrix} x_{11} \\ \vdots \\ x_{p1} \end{pmatrix}, \ldots, \begin{pmatrix} x_{1n} \\ \vdots \\ x_{pn} \end{pmatrix}.$$

Afstanden fra et punkt med koordinater $\mathbf{x} = (x_1, \ldots, x_p)'$ ned til hyperplanen 4.6 vises let at være

$$\left| \frac{\alpha_0 + \alpha_1 x_1 + \cdots + \alpha_p x_p}{\sqrt{\alpha_1^2 + \cdots + \alpha_p^2}} \right|.$$

Vi skal derfor bestemme $\alpha_0, \ldots, \alpha_p$ således, at

$$\mathrm{f}(\alpha) = \sum_{i=1}^{n} \frac{(\alpha_0 + \alpha_1 x_{1i} + \cdots + \alpha_p x_{pi})^2}{\alpha_1^2 + \cdots + \alpha_p^2}$$

minimaliseres. Indføres x_{0i} ved $x_{0i} = 1, \forall i$, kan vi skrive

$$\mathrm{f}(\alpha) = \sum_{i=1}^{n} (\sum_{j=0}^{p} \alpha_j x_{ji})^2 / \sum_{j=1}^{p} \alpha_j^2.$$

Dette minimaliseringsproblem omformuleres bekvemt til, at vi skal finde minimum af

$$\mathrm{g}(\alpha) = \sum_{i=1}^{n} (\sum_{j=0}^{p} \alpha_j x_{ji})^2$$

under bibetingelsen

$$\sum_{j=1}^{p} \alpha_j^2 = 1.$$

Indføres en Lagrange-multiplikator λ, ses, at vi skal finde det globale minimum af

$$\varphi(\alpha, \lambda) = \sum_{i=1}^{n} (\sum_{j=0}^{p} \alpha_j x_{ji})^2 - \lambda (\sum_{j=1}^{p} \alpha_j^2 - 1).$$

Vi finder for $\nu = 1, \ldots, p$

$$\frac{\partial \varphi}{\partial \alpha_\nu} = 2 \sum_{i=1}^{n} \sum_{j=0}^{p} \alpha_j x_{ji} x_{\nu i} - 2\lambda \alpha_\nu,$$

og for $\nu = 0$

$$\frac{\partial \varphi}{\partial \alpha_0} = 2 \sum_{i=1}^{n} \sum_{j=0}^{p} \alpha_j x_{ji} x_{0i} = 2 \sum_{i=1}^{n} \sum_{j=0}^{p} \alpha_j x_{ji}.$$

Vi skal sætte disse partielle afledede lig 0. Den sidste ligning bliver da

$$\sum_{i=1}^{n} \sum_{j=0}^{p} \alpha_j x_{ji} = 0,$$

eller

$$\alpha_0 = -\alpha_1 \bar{x}_{1\cdot} - \cdots - \alpha_p \bar{x}_{p\cdot}.$$

Indsættes dette i de første ligninger, kan disse omformes til

$$\sum_{i=1}^{n} \sum_{j=1}^{p} \alpha_j (x_{ji} - \bar{x}_j)(x_{\nu i} - \bar{x}_{\nu\cdot}) - \lambda \alpha_\nu = 0.$$

Kalder vi den empiriske dispersionsmatrix for observationerne for

$$\hat{\Gamma} = (\hat{\gamma}_{j\nu}),$$

ser vi, at ligningerne omformes til

$$\sum_{j=1}^{p} \alpha_j \hat{\gamma}_{j\nu} - \frac{\lambda}{n-1} \alpha_\nu = 0, \qquad \nu = 1, \ldots, p.$$

Sætter vi

$$\begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_p \end{pmatrix} = \alpha_0,$$

kan ovenstående ligningssystem skrives

$$\hat{\Gamma}\, \alpha = \frac{\lambda}{n-1} \alpha,$$

d.v.s. α er en egenvektor til $\hat{\Gamma}$ svarende til egenværdien $\lambda/n - 1$.

Spørgsmålet er nu, hvilken af de p egenværdier for $\hat{\Gamma}$ skal vi ved en konkret udregning vælge. Ved nogle manipulationer med de oprindelige ligninger (jvf. [20]p. 394) bringes man til at indse, at vi skal vælge den mindste egenværdi.

Vi har nu godtgjort sætningen. ■

BEMÆRKNING 4.10. Det resultat, der er anført i sætningen har snæver tilknytning til de resultater der gennemgås i kapitel 8 om principale komponenter. ▼

4.4.2 Ridge-regression

Ved analysen af regressionsmodeller, især modeller med mange uafhængige variable, løber man ofte ind i nogle stabilitetsproblemer. Det viser sig således ofte at udelukkelse af enkelte observationer kan forårsage voldsomme ændringer i størrelsen af de enkelte koefficienter, ja sågar bevirke et fortegnsskift. Disse problemer kan man søge undgået ved hjælp af forskellige former for trinvise regressionsprocedurer - dog ikke altid med lige stort held. I stedet for at udelukke enkelte variable helt og så focusere fuldstændigt på andre, kan man prøve at benytte lidt af den information, der ligger i alle de enkelte variable. Dette har [14] gjort med den såkaldte ridge regressionsanalyse - en metode der har vist sig at give forbedrede **prediktions** resultater.

Vi betragter den sædvanlige model

$$\mathbf{y} = \mathbf{x}\beta + \varepsilon,$$

hvor $\mathbf{x}$ er en kendt $n \times p$ matrix, β den ukendte parametervektor og ε fejlvektoren.

Vi forudsætter, at

$$\begin{aligned} \mathrm{E}(\varepsilon) &= \mathbf{0} \\ \mathrm{D}(\varepsilon) &= \sigma^2 \mathbf{I}_n. \end{aligned}$$

Den ordinære LS-estimator bliver - idet vi forudsætter at $\mathbf{x}$ har fuld rang -

$$\hat{\beta} = (\mathbf{x}'\mathbf{x})^{-1}\mathbf{x}'\mathbf{y}.$$

Vi forudsætter yderligere, at de uafhængige variable er skaleret, så $\mathbf{x}'\mathbf{x}$ er på korrelationsform (i.e. de enkelte uafhængige variable er reduceret med deres gennemsnit og divideret med deres spredning). Denne normering skal tjene til at gøre skønnene numerisk stabile, og det er altid anbefalelsesværdigt i en praktisk situation at foretage denne operation.

Hvis $\mathbf{x}'\mathbf{x}$ i denne form er nær ved en enhedsmatrix, i.e. hvis de uafhængige variable nærmest er ortogonale, er mindste kvadraters skønnet $\hat{\beta}$ udmærket. Hvis de uafhængige variable derimod er stærkt korrelerede (d.v.s. der er tale om en stor grad af **multicollinaritet**), bliver de skøn, man får frem, som ovenfor nævnt meget ustabile.

Lad os kort undersøge nogle egenskaber ved $\hat{\beta}$, som ikke er gennemgået i det tidligere. Ifølge den generelle teori haves

$$\mathrm{D}(\hat{\beta}) = \sigma^2(\mathbf{x}'\mathbf{x})^{-1}.$$

Sætter vi L lig afstanden fra $\hat{\beta}$ til β, d.v.s.

$$L^2 = (\hat{\beta} - \beta)'(\hat{\beta} - \beta),$$

fås

$$\mathrm{E}(L^2) = \sum_{i=1}^{p} \mathrm{E}[(\hat{\beta}_i - \beta_i)^2] = \sum_{i=1}^{p} \mathrm{V}(\hat{\beta}_i) = \sigma^2 \operatorname{tr}(\mathbf{x}'\mathbf{x})^{-1}.$$

Da

$$L^2 = \hat{\beta}'\hat{\beta} - 2\hat{\beta}'\beta + \beta'\beta,$$

fås, at den forventede værdi af den kvadrerede længde af $\hat{\beta}$ er

$$\mathrm{E}(\hat{\beta}'\hat{\beta}) = \sigma^2 \operatorname{tr}(\mathbf{x}'\mathbf{x})^{-1} + \beta'\beta.$$

Kaldes egenværdierne for $\mathbf{x}'\mathbf{x}$ for

$$\lambda_1 \geq \cdots \geq \lambda_p,$$

får altså (ifølge sætning 1.12 og resultatet p. 46)

$$\mathrm{E}(L^2) = \sigma^2(\frac{1}{\lambda_1} + \cdots + \frac{1}{\lambda_p}) > \frac{\sigma^2}{\lambda_p},$$

og

$$\mathrm{E}(\hat{\beta}'\hat{\beta}) = \sigma^2(\frac{1}{\lambda_1} + \cdots + \frac{1}{\lambda_p}) + \beta'\beta > \frac{\sigma^2}{\lambda_p} + \beta'\beta.$$

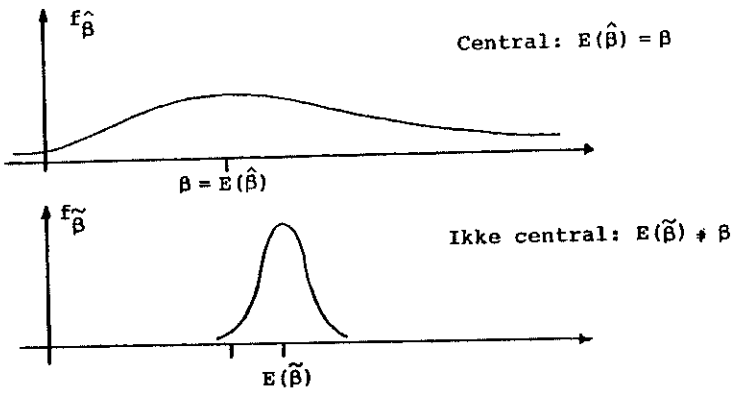
Hvis de uafhængige variable er stærkt korrelerede, vil egenværdierne for $\mathbf{x}'\mathbf{x}$ være meget forskellige, og den mindste derfor meget lille ($<< 1$). Ifølge ovenstående realtioner vil den kvadrerede afstand mellem β og $\hat{\beta}$ i dette tilfænde derfor have den står forventningsværdi, og den kvadrerede længde af $\hat{\beta}$ vil have en forventningsværdi langt større end den kvadrerede længde af β.

Denne tendens til "oppustning" af $\hat{\beta}$ skyldes kravet om centralitet. Spørgsmålet er, om man ved at undlade dette krav kan få målinger der i en vis forstand ligger "nærmere" ved β. Problemet er skitseret i figur 4.8.

Som det relevante kriterium kunne man anvende den såkaldte MSE = mean squared error. Den er (i det endimensionale tilfælde)

$$\mathrm{MSE} = \mathrm{E}[(\tilde{\beta} - \beta)^2] = \mathrm{V}(\tilde{\beta}) + \{\mathrm{E}(\tilde{\beta}) - \beta\}^2,$$

d.v.s. lig variansen på estimatoren plus kvadratet på den såkaldte **skævhed** (eng. = **bias**). Det vil åbenbart være hensigtsmæssigt at tillade en lille skævhed, hvis dette kan give en stor reduktion i variansen og dermed i MSE. Dette opnås netop med en ridge estimator.



Figur 4.8:

DEFINITION 4.2. Ved en **ridge estimator** til β i modellen

$$\mathbf{y} = \mathbf{x}\beta + \varepsilon$$

forstås en stimator $\hat{\beta}_k^* = \hat{\beta}^*$, der er løsning til

$$(\mathbf{x}'\mathbf{x} + k \cdot \mathbf{I})\hat{\beta}^* = \mathbf{x}'\mathbf{y},$$

d.v.s.

$$\hat{\beta}^* = (\mathbf{x}'\mathbf{x} + k \cdot \mathbf{I})^{-1}\mathbf{x}'\mathbf{y}.$$

Her er k en konstant $\in [0, 1]$. ▲

Vi skal nu blot citere en række egenskaber for $\hat{\beta}^*$. Disse egenskaber skal bla. bruges ved fastlæggelsen af k, der ikke er givet på forhånd.

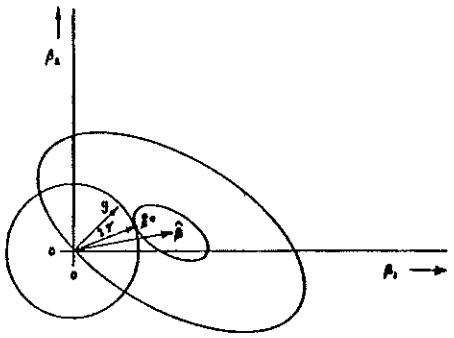
Vi har

SÆTNING 4.4. Lad situationen være som i definitionen. Vi sætter $\mathbf{x}'\mathbf{y} = g$ og betegner residualkvadratafvigelsessummen for en vilkårlig estimator $\tilde{\beta}$ for

$$H(\tilde{\beta}) = (\mathbf{y} - \mathbf{x}\tilde{\beta})'(\mathbf{y} - \mathbf{x}\tilde{\beta}).$$

Da gælder, at gradienten af H i $\tilde{\beta} = \mathbf{0}$ er proportional med og modsat rettet $\mathbf{g}$, og $\hat{\beta}_k^*$ kan fastlægges ved, at den for fastholdt længde minimaliserer $H(\tilde{\beta})$, d.v.s.

$$\min_{\|\tilde{\beta}\| = \|\hat{\beta}_k^*\|} H(\tilde{\beta}) = H(\hat{\beta}_k^*).$$



Figur 4.9:

Endvidere er $H(\hat{\beta}_k^*)$ en voksende funktion af k. Længden af $\hat{\beta}_k^*$ er en aftagende funktion af k, og vinklen γ mellem $\hat{\beta}_k^*$ og $\mathbf{g}$ er en aftagende funktion af k. ▲

BEVIS 4.5. Ikke særlig kompliceret, men forbigås. Læseren henvises til [14] og [24]. ■

Den instruktive figur 4.4.2 er taget fra [25]

Den skitserede situation geometrisk i tilfældet $p = 2$. Punktet $\hat{\beta}$ i centrum af ellipserne er LS-løsningen. Ellipserne er niveaukurver for H. Cirklen med centrum i origo er tangent til den lille ellipse. Vi bemærker, at $\hat{\beta}^*$ er den korteste vektor, der giver en residualkvadratafvigelsessum så lille som H's værdi på den lille ellipse. Endvidere bemærkes, at $\hat{\beta}^*$ altid ligger mellem $\hat{\beta}$ og $\mathbf{g}$.

Andre hovedegenskaber ved ridge skønnet fremgår af

SÆTNING 4.5. Lad situationen være som ovenfor. Da er $\hat{\beta}_k^* = \hat{\beta}^*$ en lineær transformation af $\hat{\beta}$, idet nemlig

$$\hat{\beta}^* = \mathbf{z}_k\hat{\beta} = (\mathbf{x}'\mathbf{x} + k\mathbf{I})^{-1}(\mathbf{x}'\mathbf{x})\hat{\beta}.$$

$\hat{\beta}^*$ er ikke central, idet

$$\mathrm{E}(\hat{\beta}^*) = \mathbf{z}_k\beta.$$

Dispersionsmatricen for $\hat{\beta}^*$ er

$$\mathrm{D}(\hat{\beta}^*) = \sigma^2[\mathbf{x}'\mathbf{x} + k\mathbf{I}]^{-1}(\mathbf{x}'\mathbf{x})[\mathbf{x}'\mathbf{x} + k\mathbf{I}]^{-1},$$

og den forventede kvadratiske afstand til β er

$$\mathrm{E}[(\hat{\beta}^* - \beta)'(\hat{\beta}^* - \beta)] = \mathrm{tr}(\mathrm{D}(\hat{\beta}^*)) + \beta'(\mathbf{z}_k - \mathbf{I})'(\mathbf{z}_k - \mathbf{I})\beta.$$

I det sidste udtryk er det første led lig variansen på den kvadrerede længde af $\hat{\beta}^*$, og det sidste led lig kvadratet på skævheden. ▲

BEVIS 4.6. Forbigås. Elementær øvelse i matrixmanipulationer. ■

Af sætningen følger

KOROLLAR 4.1. Hvis $\beta'\beta$ er begrænset, eksisterer der er $k > 0$, således at forventningen af den kvadratiske afstand mellem β og $\hat{\beta}^*$ er strengt mindre end forventningen af den kvadratiske afstand mellem β og $\hat{\beta}$.

BEVIS 4.7. Dette følger ved at bemærke, at $\mathrm{tr}(\mathrm{D}(\hat{\beta}^*))$ er en aftagende funktion af k, hvorimod $\beta'(\mathbf{z}_k - \mathbf{I})'(\mathbf{z}_k - \mathbf{I})\beta$ er voksende. Da $k \to 0 \Rightarrow \hat{\beta}$ ses resultatet nu umiddelbart. ■

Det eneste problem, der nu tilbagestår, er bestemmelsen af et rimeligt k. Her bruger man det såkaldte ridge trace, der begrundes med ovenstående corollar.

DEFINITION 4.3. Ved et **ridge trace** forstås en afbildning af de enkelte koefficienter i ridge estimatet som en funktion af k. ▲

Man benytter som nævnt ridge tracet ved fastlæggelsen af k. Filosofien bag dette er en følsomhedsanalyse-betragtning. Man ser på ens ridge trace, hvilke koefficienter der er følsomme over for ændringer i data. Man vælger så den mindste værdi af k, der giver et stabilt forløb af koefficienterne. Disse betragtninger vil vi anskueliggøre i

EKSEMPEL 4.2. ([25]. I eksemplet betragtes sammenhængen mellem ASTM[1] og gas chromatograf destillation af benzin. Standardmetoden at måle benzins flygtighed (i.e. den brøkdel, der er fordampet ved forskellige temperaturer) er en ASTM destillation. Gas chromatograf destillationen er langt nøjagtigere, men specifikationer for benzin er angivet i ASTM-termer. Med henblik på at kunne benytte gas chromatograf målinger til on-line kontrol af flygtigheden har man følgelig brug for en model til at forudsige ASTM destillationen af en blanding ud fra gas chromatograf destillationen.

[1] (American Standard for Testing of Materials).

I dette eksempel vil vi kun se på en enkelt ASTM temperatur, nemlig 158°F. Den afhængige variabel y angiver altså den brøkdel af benzinen, der ved en ASTM destillation er fordampet ved en temperatur på 158°F. De afhængige variable $x_1, \ldots, x_{15}$ angiver de brøkdele, der er fordampet i de respektive temperaturintervaller ved gas chromatograf metoden. Den anvendte model er

$$\mathrm{E}(y) = \beta_1 x_1 + \cdots + \beta_{15} x_{15}.$$

De uafhængige variable summerer til 1. Derfor er der ikke medtaget et konstantled.

Hovedbrugen af modellen skulle dels være forudsigelser af benzin ASTM destillationer, for hvilke forudsigelsesstandardafvigelsen skulle være $\leq 1.5\%$, og dels som input til lineær programmeringsberegninger af optimale blandeprocedurer. Tidligere analyser ved hjælp af mindste kvadraters metode og stepwise regregressionsprocedurer havde givet koefficienter, som var uacceptable fra en fysisk synsvinkel.

Der forelå 59 sammenhørende værdier af de 16 variable. Disse blev delt i to dele - en på 29 sæt sammenhørende værdier, der blev brugt til estimation af modellen, og en på 30, der blev brugt til at finde forudsigelsesfejl. I figuren p. 208 er vist ridge tracet. Man bemærker, at systemet stabiliseres for værdier omkring $k = 0.005 - 0.01$. I det aktuelle tilfælde valgtes $k = 0.006$.

I ovenstående figur vises den prediktionsstandardafvigelse, man fandt for de forskellige værdier af k. Man bemærker, at minimum indtræffer for $k = 0.006$, den fra ridge tracet valgte værdi. Mindste kvadraters analysen har en prediktions-standardafvigelse på 1.28 og ridge modellen på 1.01. I nedenstående figur er koefficienterne i de to modeller sammenlignet med koefficienter, der er resultatet af teoretiske overvejelser. Disse sidste kan selvsagt ikke anses for at være de "sande", men man forventer selvfølgelig en udpræget lighed.

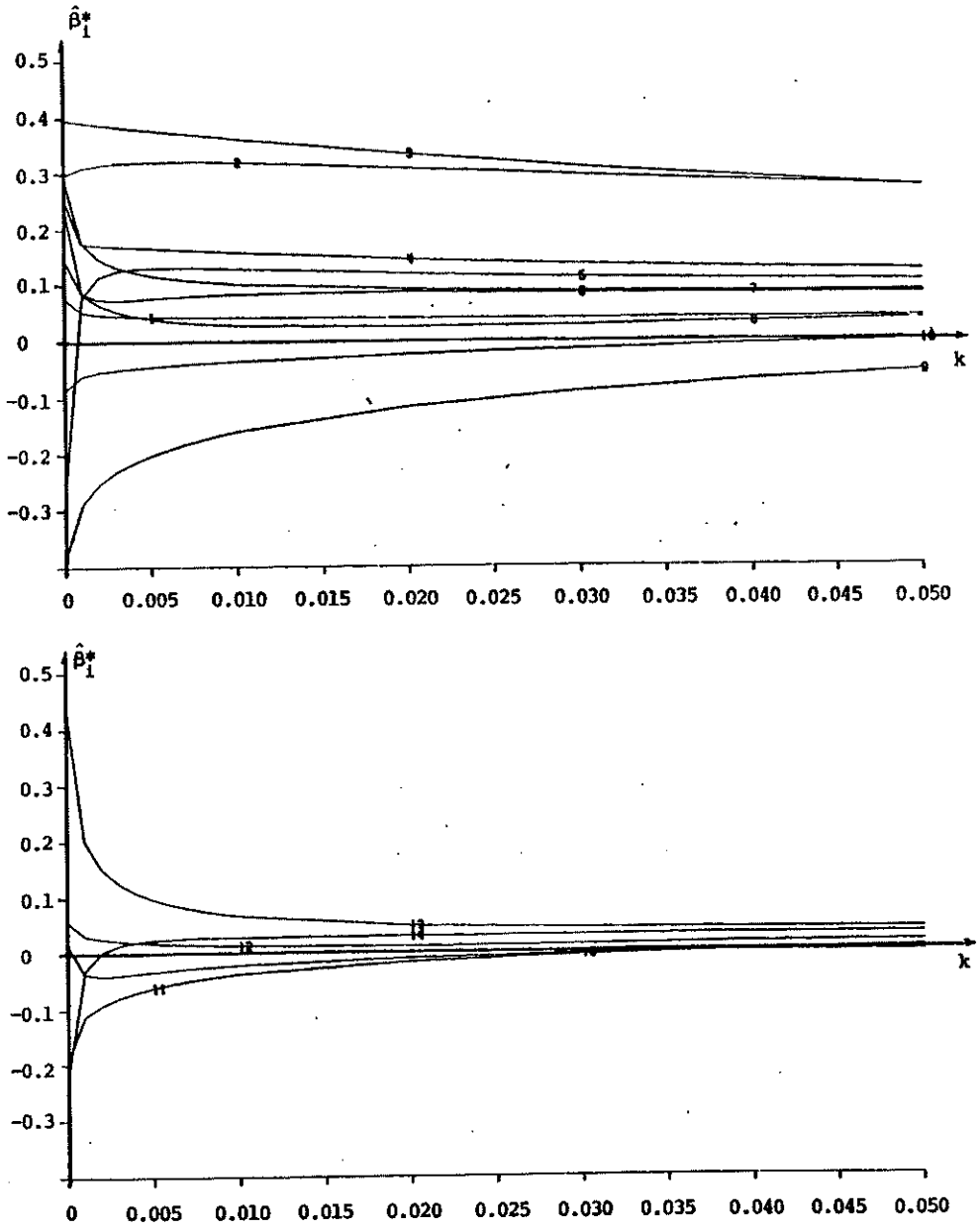
Der er en mindre uoverensstemmelse mellem den her anførte koefficient til x_1 og den i ridge tracet angivne; men det er ikke muligt ud fra kilden at afgøre, hvilken der er den korrekte.

Man ser, at ridge modellens koefficienter udviser et langt mere "roligt" forløb med stigende GC-temperaturer, og de ligger tættere ved de "teoretiske" værdier. ♦

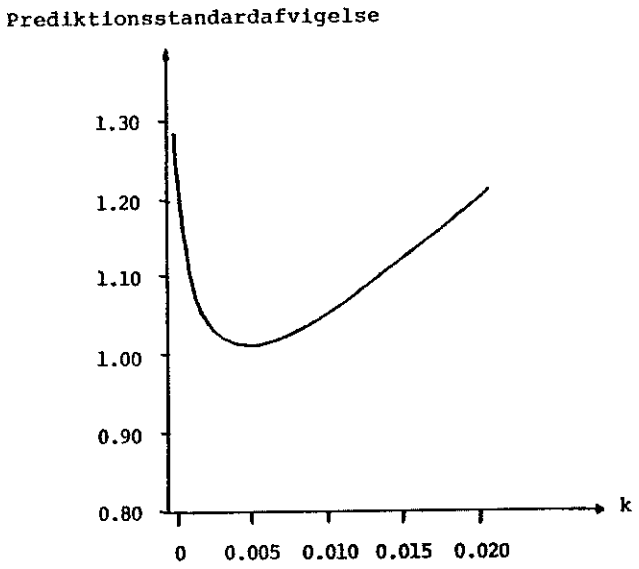
4.4.3 Ikke-lineær regression og kurvetilpasning

Ofte vil man skulle analysere regressionssituationer, der giver anledning til ikke-lineære normalligninger eller likelihoodligninger.

Man kan da selvfølgelig direkte anvende et generelt program til maximalisering af ikke-lineære funktioner eller en iterativ metode til at løse de ikke-lineære ligninger. Dispersionsmatricen for de herved fremkomne estimater kan da skønnes ved hjælp af den reciprokke informationsmatrix, jvf. afsnit 2.6.



Figur 4.10: Ridge tracet for data fra eksempel 4.2. Koefficienterne er anført i en relativ skala ($\bar{Y}$ er sat lig 0).



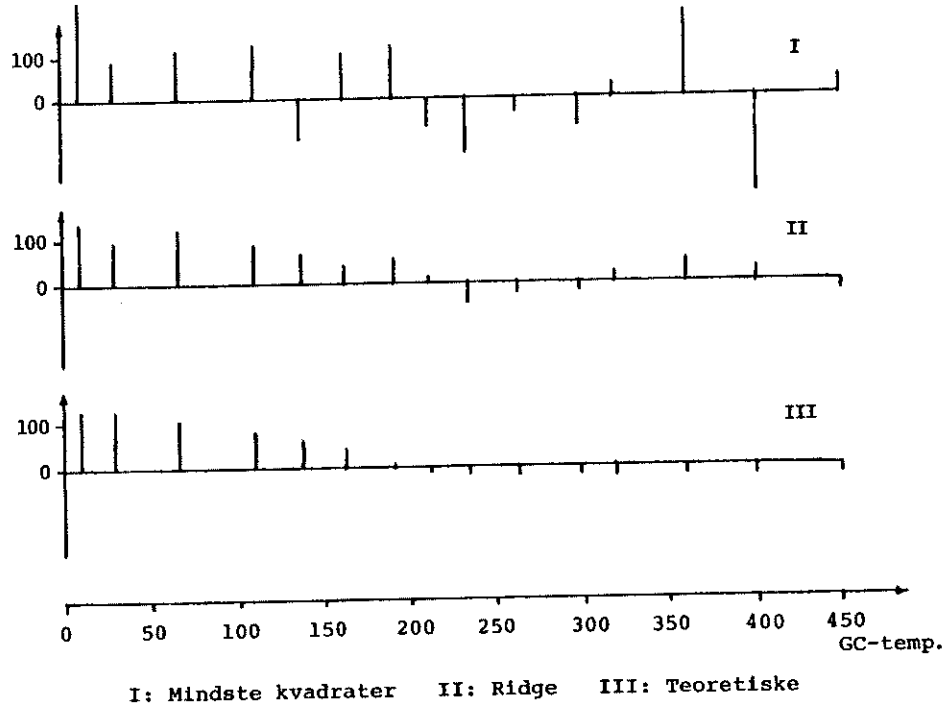
Figur 4.11: Prediktionsstandardafvigelse mod k.

Vi giver et par eksempler.

EKSEMPEL 4.3. Vi betragter nogle data, der vedrører konservering af jerngenstande fra jernalderen. Data stammer fra Eva Salomonsen (1977) [30]. På Nationalmuseets konserveringsanstalt har man i 63 år anvendt Rosenbergs glødemetode til fjernelse af klorider fra jern. For at undersøge effektiviteten af denne metode er 295 jerngenstande, konserveret i årene 1913-1974, blevet undersøgt, og antallet af defekte genstande, i.e. genstande hvor der er konstateret en fortsat nedbrydning, er opgjort for hver årgang. Tallene er anført nedenfor.

Periode	Antal undersøgte	Antal defekte	Antal defekte i % af undersøgte
1913	52	14	26.9
1921-24	34	11	32.4
1933-34	53	10	18.9
1940-43	47	13	27.7
1953-54	56	4	7.1
1961-69	46	4	8.7
1972-74	7	0	0
I alt	295	56	19.0

Antal defekte, glødede jerngenstande i forhold til det totale antal undersøgte for hver årgang.



Figur 4.12: Sammenligning af koefficienter.

Som det fremgår af tabellen, vokser defektprocenten med tiden, og denne vækst ønskes modelleret. En rimelig model vil vel være at sætte

$$\begin{aligned} X_i &= \text{antal defekte for "alder" } t_i \\ n_i &= \text{antal undersøgte for "alder" } t_i \\ p_i &= \text{sandsynligheden for, at en genstand med "alder" } t_i \text{ er defekt,} \end{aligned}$$

og så postulere, at

$$X_i \in \mathrm{B}(n_i, p_i).$$

Som "alder" vælges selvsagt den tid, der er hengået siden glødebehandlingen. For de perioder, der udstrækker sig over flere år, er glødetidspunktet sat til midten af det betragtede tidsinterval.

Det tilbagestående problem er at fastlægge defektprocenten p_t's afhængighed af tiden. Her er en meget benyttet model den logistiske kurve:

$$p_i = p(t_i) = \frac{1}{1 + \exp(-\alpha - \beta t_i)}.$$

Da kurven har asymptoter $p = 0$ og $p = 1$ og er stadigt voksende, tilfredsstiller den selvfølgelig de mest umiddelbare krav, man må stille. Defineres den såkaldte **logit**

$$\operatorname{logit} p_i = \log \frac{p_i}{1 - p_i},$$

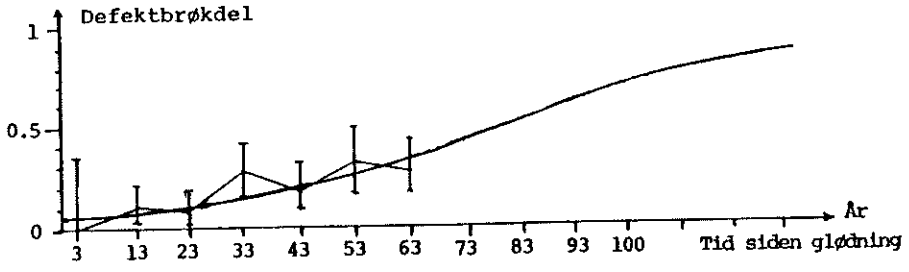
ses, at

$$\operatorname{logit} p_i = \alpha + \beta t_i,$$

d.v.s. at modellen er lineær i disse logits. Modellen er historisk især benyttet i forbindelse med bioassays, navnlig af Berkson.

Likelihoodfunktionen er

$$\mathrm{L}(\alpha, \beta) = \prod_{i=1}^{n} \binom{n_i}{x_i} \left\{ \frac{1}{1-\exp(-\alpha-\beta t_i)} \right\}^{x_i} \left\{ \frac{\exp(-\alpha-\beta t_i)}{1+\exp(-\alpha-\beta t_i)} \right\}^{n_i - x_i}$$



og dermed

$$\begin{aligned}
&\log \mathrm{L}(\alpha,\beta) = \\
&= c - \sum_i x_i \log(1+\exp(-\alpha-\beta t_i)) - \sum_i (n_i - x_i)(\alpha+\beta t_i) \\
&\quad - \sum_i (n_i - x_i)\log(1+\exp(-\alpha-\beta t_i)) \\
&= c - \sum_i n_i \log(1+\exp(-\alpha-\beta t_i)) - \sum_i (n_i - x_i)(\alpha+\beta t_i).
\end{aligned}$$

Man kan nu enten differentiere dette udtryk med hensyn til α, β og sætte differentialkoefficienterne lig 0, eller man kan maximalisere $\log \mathrm{L}(\alpha,\beta)$ ved hjælp af et generelt maximaliseringsprogram. Ved at gøre det første er fundet

$$\begin{aligned}
\hat{\alpha} &= -2.99984 \\
\hat{\beta} &= 0.03813
\end{aligned}$$

Den resulterende logistiske kurve er indtegnet i figur 4.3. Endvidere er der angivet et 95%-konfidensinterval omkring de enkelte observationer.

Der synes at være en rimelig overensstemmelse, men alligevel skal man nok være varsom med - skønt figuren ellers nok kunne friste - at foretage ekstrapolationer hundreder af år ud i fremtiden. ♦

Ofte vil man være i en situation, hvor man ønsker at tilpasse en given datamængde med en passende glat kurve; men hvor der ikke synes at være mulighed for (eller være uforholdsmæssig besværligt) at opstille en "universel" lov, der kan dække hele det betragtede område. Man kunne da tænke sig at foretage en stykkevis approximation med forskellige funktioner. Her er en meget velegnet klasse de såkaldte **spline-funktioner**, der er opkaldt efter et særligt "tegneapparat".

Disse indføres i

DEFINITION 4.4. Lad der være givet et interval $[a, b]$ og punkter $x_1, \ldots, x_n$, der alle ligger i intervallet. Lad der endvidere svare en værdi y_i til hvert x_i. Ved en **spline af orden $2m-1$** med **knuder** $x_1, \ldots, x_n$ forstås da en funktion φ, der tilfredsstiller

1) φ er et polynomium af grad $2m-1$ i $[x_i, x_{i+1}]$

2) φ er et polynomium af grad $m-1$ i $[a, x_1]$ og $[x_n, b]$

3) $\varphi, \varphi', \ldots, \varphi^{(2m-2)}$ er kontinuerte i $x_1, \ldots, x_n$

4) $\varphi(x_j) = y_j$

▲

BEMÆRKNING 4.11. En spline af orden $2m-1$ er altså sammensat "glat" af $(2m-1)$'te grads polynomier. Det kan vises, at den således fremkomne kurve meget minder om det, man ville få ved at sømme to søm ned i hver knude og dernæst tvinge en meget elastisk stålstang igennem disse (en såkaldt tegne-spline). Dette skal vi ikke komme yderligere ind på her, men blot henvise læseren til den righoldige litteratur om emnet.

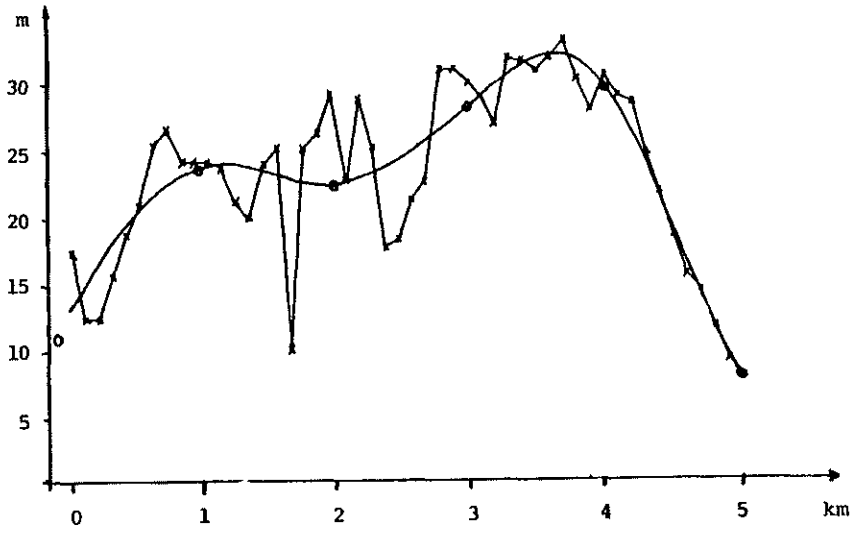
I [1] findes en række programmer til at approximere givne data med kubiske spline-funktioner, dels på basis af på forhånd specificerede knuder og dels ved "selv" at vælge knuder. ▼

EKSEMPEL 4.4. I nedenstående figur er anført koten for en række punkter på en linie af længden 5 km i Dyrehaven. Data er stillet til rådighed af Poul Frederiksen, Institut for Landmåling og Fotogrammetri. Det er ved forskellige projekter interessant at beregne et udtryk for en variation omkring en passende valgt "glat" "trendkurve" (trend = tendens). Det forekommer derfor oplagt at indlægge en kubisk spline-funktion. Dette er gjort ved hjælp af Harwell programmet VBO5B, der på basis af observationer $(x_1, y_1), \ldots, (x_m, y_m)$ minimaliserer en størrelse

$$F = \sum_{i=1}^{m} w_i^2 \{y_i - S(x_i)\}^2,$$

hvor w_i'er er brugerspecificerede vægte, og S er en kubisk spline-funktion med knuder $j, j = 1, \ldots, n$, hvis abscisser er specificeret af brugeren.

Den resulterende spline og dens knuder er ligeledes anført. Man bemærker den overordentlig nydelige tilpasning mellem de meget uregelmæssige observationer og spline-funktionen. ♦



Figur 4.13: Koter og approximerede kubiske spline.

Kapitel 5
Variansanalyser

5.1 Indledning

Variansanalyser er en fællesbetegnelse for en lang række statistiske metoder, der bygger på en model, der kan udtrykkes:

Observeret værdi	$= \sum_i$	(parametre, som repræsenterer anviselige effekter)
	$+ \sum_i$	(stokastiske variable, som repræsenterer anviselige effekter)
	$+$	stokastisk variabel, som repræsenterer residuale effekter.

Udtrykt formelt:

$$X = \alpha_1 + \cdots + \alpha_m + A_1 + \cdots + A_n + Z.$$

Her betyder anviselig effekt en effekt, der skyldes faktorer, som vi tager hensyn til i vort eksperiment. Hvis man e.g. forestiller sig, at den observerede værdi X er udbyttet af en kemisk reaktion, vil tilstedeværelsen af en bestemt katalysator være en **faktor**, som vi registrerer, og som vi regner med, har en indflydelse på udbyttet. Det vil formentligt være mest rimeligt at regne denne effekt som en konstant. Hvis udbyttet også afhænger af de meteorologiske forhold, vil det være rimeligt at knytte en stokastisk effekt til udtrykket. Ved at fortsætte med at indføre effekter er det oftest muligt at få fjernet det meste af residualvariationen. Man må derfor vælge at standse "indtagningen" af faktorer, når det skønnes, at residualvariationen ikke mindskes tilstrækkeligt herved.

Vi forudsætter

1. at forventningsværdien af hver residualvariabel Z er 0,

2. at de er stokastisk uafhængige,

3. at de har samme varians, og

4. at de er normalt fordelte.

Af disse betingelser er 3. den mest afgørende, idet analyserne af resultaterne hele tiden sker ved at måle forskelle mellem middelværdier relativt til residualvariansen (eller rettere spredningen).

I de tilfælde, hvor vi kun har determiniske komponenter $\alpha_1, \ldots, \alpha_m$, er det klart, at vi har en generel, lineær model, og alle resultater om estimatorer og testning kan umiddelbart afledes som specialtilfælde af det tidligere gennemgåede. I de øvrige tilfælde må vi udlede resultaterne særskilt. Af tidsmæssige grunde vil vi dog ofte kun anføre en heuristisk begrundelse for de postulerede resultater.

Hvis alle A_i'erne er 0, d.v.s., hvis vi har modellen

$$X = \alpha_i + \ldots + \alpha_m + Z,$$

taler vi om en model med **systematisk variation**. I modellen

$$X = (\mu+)A_1 + \ldots + A_n + Z,$$

(hvor μ er et niveaumål) taler vi om **tilfældig variation**. De øvrige kaldes **blandede modeller**. Man taler også om **type I** og **type II** variansanalyser. Type II-modellen (tilfældig variation) kaldes også en **varianskomponentmodel** og type I-modellen (systematisk variation) en ren **parametrisk model**.

Vi går nu over til at betragte den simpleste model, hvor vi kun har én faktor.

5.2 Ensidet variansanalyse

En del af det følgende stof vil være en repetition af det i bind 1 gennemgåede om den ensidede variansanalyse, men dette er gjort for at få en sammenhængende beskrivelse af teorien.

5.2.1 Modeller

Vi har givet observationerne

$$\begin{array}{lll} X_{11}, & \ldots, & X_{1n_1} \\ \vdots & & \\ X_{k1}, & \ldots, & X_{kn_k} \end{array}$$

svarende til k niveauer af en enkelt faktor. (Hvis man e.g. måler temperaturen i noget smeltet tin med 5 forskellige termometre, har vi én faktor, nemlig termometer, og denne har 5 "niveauer").

Vi har de mulige modeller

$$\begin{array}{lllll} I) & X_{ij} = \mu + \alpha_i + Z_{ij}, & i = 1, \ldots, k, & j = 1, \ldots, n_i \\ II) & X_{ij} = \mu + A_i + Z_{ij}, & i = 1, \ldots, k, & j = 1, \ldots, n_i \end{array}$$

Her betegner μ et generelt niveau, og α_i henholdsvis A_i angiver effekten af, at vor måling er foretaget med faktoren i niveau i.

Vi forudsætter

$$I) \quad \sum_i n_i \alpha_i = 0$$
$$Z_{ij} \in \mathrm{N}(0, \sigma^2), \qquad \text{stokastisk uafhængige for alle } i, j.$$

$$II) \quad \left.\begin{array}{ll} A_i & \in \mathrm{N}(0, \sigma_a^2) \\ Z_{ij} & \in \mathrm{N}(0, \sigma^2) \end{array}\right\} \quad \text{stokastisk uafhængige for alle } i, j.$$

Vi vil nu kort belyse forskellen på de to modeller.

I tilfælde I har vi, at

$$\mathrm{Cov}(X_{i\nu}, X_{i\mu}) = \mathrm{Cov}(Z_{i\nu}, Z_{i\mu}) = 0, \quad \mu \neq \nu,$$

og derfor er observationerne i samme gruppe ukorrelerede og dermed uafhængige (grundet normaliteten). I tilfælde II gælder derimod, stadig for $\mu \neq \nu$,

$$\begin{aligned} \mathrm{Cov}(X_{i\nu}, X_{i\mu}) &= \mathrm{Cov}(A_i + Z_{i\nu}, A_i + Z_{i\mu}) \\ &= \mathrm{Cov}(A_i, A_i) \\ &= \sigma_a^2, \end{aligned}$$

d.v.s. målinger fra samme gruppe er **ikke** uafhængige.

Ser vi alene på den enkelte observation gælder

$$\begin{array}{ll} I) & \mathrm{V}(X_{ij}) = \mathrm{V}(Z_{ij}) = \sigma^2 \\ II) & \mathrm{V}(X_{ij}) = \mathrm{V}(A_i) + \mathrm{V}(Z_{ij}) = \sigma^2 + \sigma_a^2. \end{array}$$

Vi har altså, at variansen i en type II model er lig summen af residualvariansen σ^2 og variansen på den tilfældige komponent A_i. Heraf fremgår også, hvorfor type II-modellen kaldes en varianskomponentmodel.

Den væsentlige forskel mellem de to modeller er, at vi i den ene situation **vælger** at tolke forskelle som udtryk for stokastiske. Hvilket valg, der er det rimelige, afhænger selvfølgelig af forsøgsomstændighederne og af, hvad resultatet skal bruges til, i.e. af, hvorledes en gentagelse af experimentet ville være. Dette forhold vil vi belyse nedenfor i et eksempel.

Først vil vi dog lige præcisere det "tekniske" mål for den statistiske analyse. Det vil i første omgang være: at estimere de involverede parametre, i.e.

$$\alpha_1, \ldots, \alpha_k,$$

respektive

$$\sigma_a^2,$$

og i anden omgang: at teste hypoteserne

$$H_I : \alpha_1 = \cdots = \alpha_k = 0 \quad \text{mod} \quad K_I : \exists\, \alpha_i \neq 0,$$

respektive

$$H_{II} : \sigma_a^2 = 0 \quad \text{mod} \quad K_{II} : \sigma_a^2 > 0.$$

Accept af disse tests giver begge, at niveauet af faktoren ingen betydning har. Den eneste variation, der er i materialet, er residualvariationen.

Lad os søge at klargøre forskellen på de to modeller ved hjælp af et

EKSEMPEL 5.1. I nedenstående tabel er anført resultatet af en række målinger af trækstyrke af noget gummi (målt i pounds per square inch).

A	B	C	D
3210	3225	3220	3545
3000	3320	3410	3600
3315	3165	3320	3580
	3145	3370	3485

Hvis betegnelserne A, B, C, D angiver, hvilken af 4 forskellige produktionsmetoder, der er anvendt ved fremstillingen af de gummitråde, der danner basis for de pågældende målinger, vil det være mest rimeligt at anvende den systematiske model

$$X_{ij} = \mu + \alpha_i + Z_{ij},$$

hvor α_i angiver effekten af den i'te metode.

Hvis vi derimod blot har udtaget 4 forskellige kasser fra en og samme fabriks lager, og dernæst målt styrken af gummi fra disse kasser med henblik på at få et skøn over styrken af fremtidige gummiprøver, vil det være mere rimeligt at anvende modellen

$$X_{ij} = \mu + A_i + Z_{ij},$$

hvor A_i altså angiver den **tilfældige** afvigelse, vi har fra middelkvaliteten i denne tilfældigt valgte kasse. Hvis vi her havde anvendt en systematisk model, kunne vi (formelt) estimere et α_i. Et sådant α_i vil imidlertid ikke være af nogen synderlig interesse. Hvis man en anden dag tilfældigt udvælger sig en kasse, har man ikke fået nogen oplysning om kvaliteten i denne kasse, selv om man kender $\alpha_1, \ldots, \alpha_4$. Hvis man derimod anvender model II og anfører et estimat for σ_a^2, kan vi regne med, at kvaliteten af gummiet i den foreliggende kasse "nok er μ + et tilfældigt bidrag fra en $N(0, \sigma_a^2)$-fordeling" (plus et tilfældigt bidrag fra residualen). (Vi må lige præcisere, at det i den systematiske model selvfølgelig er overordentligt interessant at få estimeret parametrene α_i, idet vi jo så kan angiver et skøn over trækstyrken for gummi, produceret efter de 4 forskellige metoder).

Vi ser her et eksempel på, hvor det er afgørende at gøre sig klar hvad en **gentagelse af eksperimentet** vil være. Hvis en gentagelse her i tilfælde 2 blot ville være tilfældigt at udvælge gummiprøver fra en af de allerede udvalgte 4 kasser, ja da vil vi selvfølgelig skulle anvende den systematiske model. det er jo da interessant om vi vælger fra en kasse, som vi på forhånd har fundet er "god", eller om vi vælger fra en kasse som er "dårlig". Den en gang erkendte forskel mellem de fire kasser vil vi selvfølgelig ikke se bort fra, og blot kalde tilfældig, hvis vi skal udtage prøver fra de selvsamme kasser. ♦

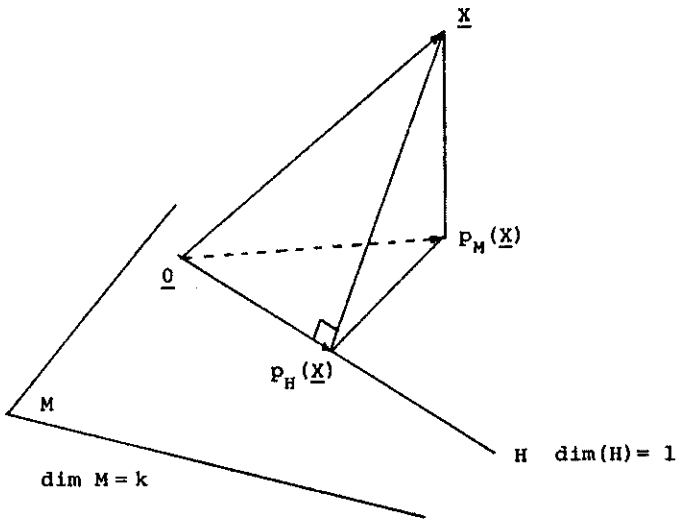
5.2.2 Analyse af den systematiske model

Fra det elementære kursus vides, at variansanalyseskemaet bliver (hypotese $\alpha_1 = \cdots = \alpha_k = 0$)

Variation	SAK	Fr. gr.
Hypotese-model	$\sum_i n_i(\bar{X}_i - \bar{X})^2$	$k-1$
Model-observationer	$\sum_i \sum_j (X_{ij} - \bar{X}_i)^2$	$N-k$
Hypotese-observationer	$\sum_i \sum_j (X_{ij} - \bar{X})^2$	$N-1$

hvor

$$
\begin{aligned}
N &= \sum_i n_i \\
\bar{X}_i &= \frac{1}{n_i} \sum_j X_{ij} \\
\bar{X} &\quad \frac{1}{N} \sum_i \sum_j X_{ij}.
\end{aligned}
$$



Udtrykt i sproget fra den generelle model har vi, at M er det underrum af R^N, hvor de n_1 første koordinater er ens, de n_2 næste koordinater er ens,. . ., de n_k sidste koordinater er ens. H er det underrum af M, hvor samtlige koordinater er ens. Det er klart, at

dim $M = k$ og dim $H = 1$. Vi har da med den sædvanlige norm

$$\|\mathbf{X} - p_M(\mathbf{X})\|^2 = \sum_i \sum_j (X_{ij} - \bar{X}_i)^2$$
$$\|p_H(\mathbf{X}) - p_M(\mathbf{X})\|^2 = \sum_i n_i(\bar{X}_i - \bar{X})^2.$$

Spaltningssætningen giver umiddelbart, at disse er stokastisk uafhængige og $\sigma^2\chi^2(f)$-fordelte under H_0. Hermed er den i bind 1 postulerede uafhængighed af disse kvadratafvigelsessummer godtgjort. Vi vil nu bestemme forventningsværdierne af disse størrelser - også uden betingelsen, at H_0 er rigtig.

For nemheds skyld anvender vi betegnelserne

$$\text{SAK}_1 = \|\mathbf{X} - p_M(\mathbf{X})\|^2 = \sum_i \sum_j (X_{ij} - \bar{X}_i)^2,$$

og

$$\text{SAK}_2 = \|p_H(\mathbf{X}) - p_M(\mathbf{X})\|^2 = \sum_i n_i(\bar{X}_i - \bar{X})^2.$$

Vi har altid, at

$$\text{SAK}_1 \in \sigma^2\chi^2(N - k).$$

Da

$$\begin{aligned} \text{SAK}_2 &= \sum_i n_i(\bar{X}_i - \bar{X})^2 = \sum_i n_i(\bar{Z}_i - \bar{Z} + \alpha_i)^2 \\ &= \sum_i n_i(\bar{Z}_i - \bar{Z})^2 + \sum_i n_i\alpha_i^2 + 2\sum_i n_i\alpha_i(\bar{Z}_i - \bar{Z}), \end{aligned}$$

er

$$\begin{aligned} \text{E}(\text{SAK}_2) &= \text{E}(\sum_i n_i(\bar{Z}_i - \bar{Z})^2) + \sum_i n_i\alpha_i^2 + 2\sum_i n_i\alpha_i 0 \\ &= \sigma^2(k-1) + \sum_i n_i\alpha_i^2. \end{aligned}$$

Leddet $\sigma^2(k-1)$ skyldes, at størrelserne Z_{ij} er $\in \text{N}(0, \sigma^2)$ og derfor tilfredsstiller hypotesen om fuldstændig homogenitet. En ensidet variansanalyse på Z_{ij}'erne giver da (v.h.a. spaltningssætningen)

$$\sum n_i(\bar{Z}_i - \bar{Z})^2 \in \sigma^2\chi^2(k-1),$$

og heraf følger resultatet. Vi kan derfor supplere variansanalyseskemaet p. 220 med en søjle med de forventede værdier af SAK_i/f_i'erne:

$S^2 = \frac{\mathrm{SAK}}{f}$	$\mathrm{E}(S^2)$
$S_2^2 = \frac{\mathrm{SAK}_2}{f_2}$	$\sigma^2 + \frac{1}{k-1}\sum n_i\alpha_i^2$
$S_1^2 = \frac{\mathrm{SAK}_1}{f_1}$	σ^2

Vi bemærker, at disse middelværdier smukt anskueliggør, hvorfor et test af hypotesen

$$H_I : \alpha_1 = \cdots = \alpha_k = 0 \quad \text{mod} \quad K_I : \exists i : \alpha_i \neq 0.$$

må have et kritisk område af formen

$$\{\mathbf{x} \left| \frac{s_2^2}{s_1^2} > c \right.\}.$$

5.2.3 Analyse af den tilfældige model

Vi er her som nævnt interesseret i test af hypotesen

$$H_0 : \sigma_a^2 = 0 \quad \text{mod} \quad H_1 : \sigma_a^2 > 0.$$

Hvis hypotesen er sand, er A_i lig 0 med sandsynligheden 1, d.v.s. vi har, at

$$X_{ij} = \mu + Z_{ij},$$

hvor $Z_{ij} \in \mathrm{N}(0, \sigma^2)$. Der er da ingen særlig gruppeeffekt - resultaterne afviger kun fra hinanden ved den tilfældige fejl Z_{ij}. Vi analyserer nu de samme kvadratafvigelsessummer, som vi havde i den systematiske model. Vi har

$$\begin{aligned} \mathrm{SAK}_1 &= \sum_i \sum_j (X_{ij} - \bar{X}_i)^2 \\ &= \sum_i \sum_j (Z_{ij} - \bar{Z}_i)^2 \in \sigma^2\chi^2(N-k). \end{aligned}$$

Under H_0 er

$$\begin{aligned} \mathrm{SAK}_2 &= \sum_i n_i(\bar{X}_i - \bar{X})^2 \\ &= \sum_i n_i(\bar{Z}_i - \bar{Z})^2 \in \sigma^2\chi^2(k-1), \end{aligned}$$

og denne størrelse er **uafhængig** af SAK_1 (brug e.g. spaltningssætningen på de variable Z_{ij}). Generelt har vi

$$\begin{aligned}\mathrm{SAK}_2 &= \sum n_i(\bar{X}_i - \bar{X})^2 \\ &= \sum n_i(\bar{Z}_i - \bar{Z} + A_i - \bar{A})^2 \\ &= \sum n_i(\bar{Z}_i - \bar{Z})^2 + \sum n_i(A_i - \bar{A})^2 \\ &\quad +2\sum n_i(\bar{Z}_i - \bar{Z})(A_i - \bar{A}),\end{aligned}$$

hvor

$$\bar{A} = \frac{1}{N}\sum_i n_i A_i.$$

Heraf fås

$$\mathrm{E}(\mathrm{SAK}_2) = \sigma^2(k-1) + \mathrm{E}(\sum_i n_i(A_i - \bar{A})^2 + 0).$$

Nu er

$$\sum_i n_i(A_i - \bar{A})^2 = \sum n_i A_i^2 - N\bar{A}^2,$$

og dermed

$$\begin{aligned}\mathrm{E}(\sum_i n_i(A_i - \bar{A})^2) &= \sum n_i\,\mathrm{E}(A_i^2) - N\,\mathrm{E}(\bar{A}^2) \\ &= \sum n_i\sigma_a^2 - N\frac{1}{N^2}\sum n_i^2\sigma_a^2 \\ &= \sigma_a^2(N - \frac{1}{N}\sum n_i^2).\end{aligned}$$

Det andet lighedstegn følger let, når man erindrer, at $\mathrm{E}(A_i^2) = \mathrm{V}(A_i)$ og $\mathrm{E}(\bar{A}^2) = \mathrm{V}(\bar{A})$, da disse stokastiske variable har forventningsværdierne 0.

Samles ovenstående, fås

$$\mathrm{E}(\mathrm{SAK}_2/(k-1)) = \sigma^2 + \frac{1}{k-1}(N - \frac{1}{N}\sum n_i^2)\sigma_a^2.$$

Er alle n_i'er lig n, fås specielt

$$\mathrm{E}(S_2^2) = \mathrm{E}(\mathrm{SAK}_2/(k-1)) = \sigma^2 + n\sigma_a^2.$$

Da $\mathrm{E}(S_1^2) = \sigma^2$, og da $S_2^2/S_1^2 \in \mathrm{F}(k-1, N-k)$ under H_0, ses, at også i den tilfældige model er det kritiske område af formen

$$\{\mathbf{x} \left| \frac{s_2^2}{s_1^2} > \mathrm{F}(k-1, N-k)_{1-\alpha} \right.\},$$

denne gang er det blot hypotesen $\sigma_a^2 = 0$, vi undersøger.

Vi har altså fået den samme teststørrelse og det samme kritiske område som under analysen af den parametriske model. Dette kunne måske forlede læseren til at tro, at de to situationer i virkeligheden ikke adskiller sig fra hinanden. At dette ikke er tilfældet, kan bl.a. ses af, at teststørrelsernes fordeling **uden** H_0 ikke er ens.

I det parametriske tilfælde er

$$\frac{s_2^2}{s_1^2} \in \mathrm{F}(k-1, N-k; \frac{1}{\sigma^2}\sum n_i\alpha_i^2),$$

altså ikke-centralt F-fordelt. Dette resultat er, om end en ikke helt triviel følge af, så dog ikke uforståeligt ved sammenligning med definitionen på ikke-centrale χ^2- og F-fordelinger.

Hvis vi indskrænker os til det **balancerede** tilfælde (i.e. hvor alle n_i'er er ens), har vi i den tilfældige model

$$\begin{aligned} \mathrm{SAK}_2 &= n\sum_i[(\bar{Z}_i + A_i) - (\bar{Z} + \bar{A})]^2 \\ &= n\sum_i(V_i - \bar{V})^2, \end{aligned}$$

hvor

$$V_i = (\bar{Z}_i + A_i) \in \mathrm{N}(0, \frac{1}{n}\sigma^2 + \sigma_a^2).$$

Derfor er

$$\mathrm{SAK}_2 \in (\sigma^2 + n\sigma_a^2)\chi^2(k-1).$$

Da det endvidere indses, at SAK_2 og SAK_1 fremdeles er uafhængige, har vi altså

$$\begin{aligned}\frac{S_2^2}{S_1^2} &\in \frac{1}{\sigma^2}(\sigma^2 + n\sigma_a^2)\mathrm{F}(k-1, N-k) \\ &= (1 + n\sigma_a^2/\sigma^2)\mathrm{F}(k-1, N-k),\end{aligned}$$

d.v.s. **centralt** F-fordelt med skalaparameter $1 + n\sigma_a^2/\sigma^2$, d.v.s. et helt andet resultat, end vi havde i den parametriske model.

5.2.4 Resumé af analyserne og et eksempel

Vi opstiller nu et variansanalyseskema med de forventede værdier af S^2-størrelserne indsat.

Variation	SAK	D.F.	S^2	$\mathrm{E}(S^2)$ Systematisk model	$\mathrm{E}(S^2)$ tilfældig model
Mellem grupper	SAK_2	$k-1$	S_2^2	$\sigma^2 + \frac{1}{k-1}\sum_i n_i\alpha_i^2$	$\sigma^2 + \frac{1}{k-1}[N - \frac{1}{N}\sum_i n_i^2]\sigma_a^2$
Inden f. grupper	SAK_1	$N-k$	S_1^2	σ^2	σ^2
Total	SAK_0	$N-1$			

hvor

$$\begin{aligned}\mathrm{SAK}_2 &= \sum_i n_i(\bar{X}_i - \bar{X})^2 \\ \mathrm{SAK}_1 &= \sum_i \sum_j (X_{ij} - \bar{X}_i)^2 \\ \mathrm{SAK}_0 &= \sum_i \sum_j (X_{ij} - \bar{X}_i)^2\end{aligned}$$

Det ses, at et centralt skøn over σ_a^2 er

$$\hat{\sigma}_a^2 = (S_2^2 - S_1^2)/(\frac{1}{k-1}(N - \frac{1}{N}\sum n_1^2)).$$

Dette skøn kan antage negative værdier, skønt det er et estimat over en ikke-negativ størrelse, nemlig σ_a^2. Det er i sådanne tilfælde sædvane at sætte $\hat{\sigma}_a^2 = 0$.

De kritiske områder for test af

$$\alpha_1 = \cdots = \alpha_k = 0$$

mod alle alternativer, respektive for test af

$$\sigma_a^2 = 0 \quad \text{mod} \quad \sigma_a^2 > 0,$$

er ens og lig

$$\{(x_{11}, \ldots, x_{kn_k}) | s_2^2/s_1^2 > \mathrm{F}(k-1, N-k)_{1-\alpha}\}$$

ved test på niveau α. Fordelingerne af S_2^2/S_1^2 i de to tilfælde er forskellige, når H_0 ej er opfyldt.

EKSEMPEL 5.2. Vi betragter de p. 218 anførte data. Af hensyn til beregningerne koder vi data v.h.a. relationen

$$x'_{ij} = \frac{x_{ij} - 3000}{5}.$$

Vi udfører nu de sædvanlige beregninger til en ensidet variansanalyse og får variansanalyseskemaet

Variation	SAK	Fr. gr.	S^2	$E(S^2)$ Systematisk model	$E(S^2)$ tilfældig model
Mellem grup.	12961.98	3	4320.66	$(\sigma^2 + \frac{1}{3}\sum n_i\alpha_i^2)\frac{1}{25}$	$(\sigma^2 + 3.73\sigma_a^2)\frac{1}{25}$
inden f. grup.	3911.75	11	355.61		
Total	16873.73	14			

Da

$$\frac{4320.66}{355.61} = 12.15 \simeq \mathrm{F}(3, 11)_{0.9991},$$

forkastes hypotesen

$$\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0.$$

(hvis vi arbejder med den parametriske model), henholdsvis hypotesen

$$\sigma_a^2 = 0$$

(hvis vi arbejder med den tilfældige model).

Arbejder vi med den systematiske model $X_{ij} = \mu + \alpha_i + Z_{ij}$, må vi konkludere, at det ikke kan antages, at metoderne er ens. Vi har estimaterne

$$\begin{aligned} \hat{\mu} &= 3327.33 \\ \hat{\alpha}_1 &= -152.33 \\ \hat{\alpha}_2 &= -113.58 \\ \hat{\alpha}_3 &= 2.67 \\ \hat{\alpha}_4 &= 225.17 \end{aligned}$$

og

$$\hat{\sigma}^2 = 25 \cdot 355.61 = 94.28^2.$$

Arbejder vi med den tilfældige model $X_{ij} = \mu + A_{ij} + Z_{ij}$, har vi, at vi må antage at der er tilfældige fluktuationer imellem de forskellige kasser. Disse udtrykkes ved estimaterne

$$\begin{aligned} \hat{\mu} &= 3327.33 \\ \hat{\sigma}^2 &= 25 \cdot 355.61 = 94.28^2 \\ \hat{\sigma}_a^2 &= 26551.67 = 162.95^2 \end{aligned}$$

♦

Vi har i dette afsnit gjort grundigt rede for, hvorledes sammenhængen er mellem analysen af en tilfældig model og en tilsvarende systematisk. Vi vil derfor i de følgende afsnit nøjes med at præcisere de relevante resultater uden at give et bevis for disse.

5.3 Tosidet variansanalyse. Hierarkisk klassifikation og krydsklassifikation

5.3.1 Hierarkisk klassifikation og krydsklassifikation

Vi skal nu beskæftige os med analysen af forsøg, hvor man undersøger virkningen af 2 faktorer A og B. Alt efter formålet med en sådan undersøgelse får vi 2 typer af modeller, nemlig den såkaldte **krydsklassifikation** og den såkaldte **hierarkiske klassifikation**. Lad os give et lille eksempel, inden vi anfører de stringente definitioner.

EKSEMPEL 5.3. Ved den afregning, en kornproducent modtager for noget leveret korn, indgår vandindholdet i % som en meget afgørende faktor. Lad os antage, at en bonde har 3 siloer med korn, som han ønsker at sælge. Kornhandleren tager e.g. 3

stikprøver fra 4 tilfældigt valgte steder i hver silo (ikke nødvendigvis de samme steder i hver silo) og måler vandindholdet i kornet. Vi har altså målinger

$$\begin{array}{lllll} X_{ij\nu}, & i & = & 1,2,3 & \text{(silo nr.)} \\ & j & = & 1,2,3,4 & \text{(målested nr.)} \\ & \nu & = & 1,2,3 & \text{(gentagelses nr.)} \end{array} .$$

Som en model til beskrivelse af måleresultaterne kan vi vælge

$$X_{ij\nu} = \mu + \alpha_i + T_{j(i)} + Z_{\nu(ij)},$$

$i = 1,2,3$ (silonr.), $j = 1,2,3,4$ (målestednr.), $\nu = 1,2,3$ (gentagelsesnr.). Her betegner μ et generelt niveau, α_i den i'te silos afvigelse fra niveauet og $T_{j(i)}$ den (tilfældige) afvigelse fra niveauet i den i'te silo, som der er det j'te målested i denne silo. $Z_{\nu(ij)}$ er residualet. Vi skal senere kommentere den specielle anvendelse af parenteser ved indiceringen.

Lad os nu betragte en anden situation, hvor bonden er interesseret i, hvorledes vandindholdet i kornet fordeler sig på 4 specielt valgte steder i siloerne (e.g. ved bunden, $\frac{1}{3}$ oppe, $\frac{2}{3}$ oppe og ved toppen). Hvis han derfor lader udtage 3 prøver fra hvert af stederne i hver silo og måler vandindholdet, har vi igen målinger

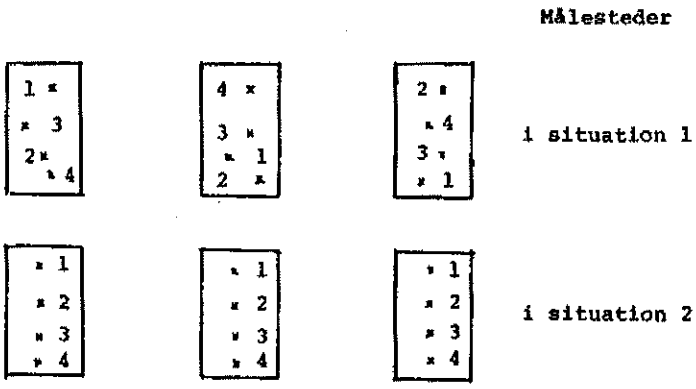
$$\begin{array}{lllll} X_{ij\nu}, & i & = & 1,2,3 & \\ & j & = & 1,2,3,4 & \text{og} \\ & \nu & = & 1,2,3 & \end{array} .$$

Her vil vi vælge modellen

$$X_{ij\nu} = \mu + \alpha_i + \beta_j + \theta_{ij} + Z_{\nu(ij)}.$$

Her er μ og α_i fremdeles niveau og den i,te silos afvigelse fra dette. β_j angiver målestedets afvigelse fra niveauet og θ_{ij} vekselvirkningen mellem målested og silo. $Z_{\nu(ij)}$ er residualet.

Vi ser, at det ikke ville være rimeligt at anvende denne model i det første tilfælde; thi dér var der ingen sammenhæng mellem målestederne i de forskellige siloer. Der havde de tre prøver fra målested nr. 1 i silo 1 intet til fælles med de tre prøver fra målested nr. 1 i silo 2 (andet end, at de tilfældigvis er taget som de første i hver silo; det første målested kan derimod være i bunden i den ene silo i midten i den anden etc.). I den anden situation stammer de begge fra e.g. toppen af de respektive siloer, og det vil derfor være rimeligt at indføre en stedparameter β_j. Det er disse parametre, som undersøgelse nr. 2 sigter mod at få belyst (estimeret), således at man kan komme med udsagn som, at kornet i toppen af en silo indeholder gennemsnitligt $\frac{1}{2}\%$ mindre vand end det ved bunden af siloen etc. I undersøgelse nr. 1 er det af hovedinteresse at få estimeret μ og α_i, således at man kan udtale sig om det gennemsnitlige vandindhold i



Figur 5.1:

hver af de 3 siloer. Hvis man er interesseret i, hvor meget vandindholdet varierer i de enkelte siloer, kan man estimere $\mathrm{V}(T_{j(i)})$.

Vi ser altså, at vi med de $3 \times 4 \times 3$ målinger har 2 mulige modeller. Hvilken der skal anvendes, afhænger som sagt af, på hvilken måde data er indsamlet og med hvilket formål. ♦

For at kunne skelne de to situationer fra hinanden vælger vi den anførte skrivemåde med at anbringe i i en parentes efter j, hvis j er underordnet i d.v.s. hvis det er sådan, at sted j i den i'te silo intet har tilfælles med sted j i den i'te silo. Analogt gælder selvfølgelig for en indicering som $\nu(ij)$ etc.

Efter dette eksempel er vi nu i stand til at formulere de to typer af modeller.

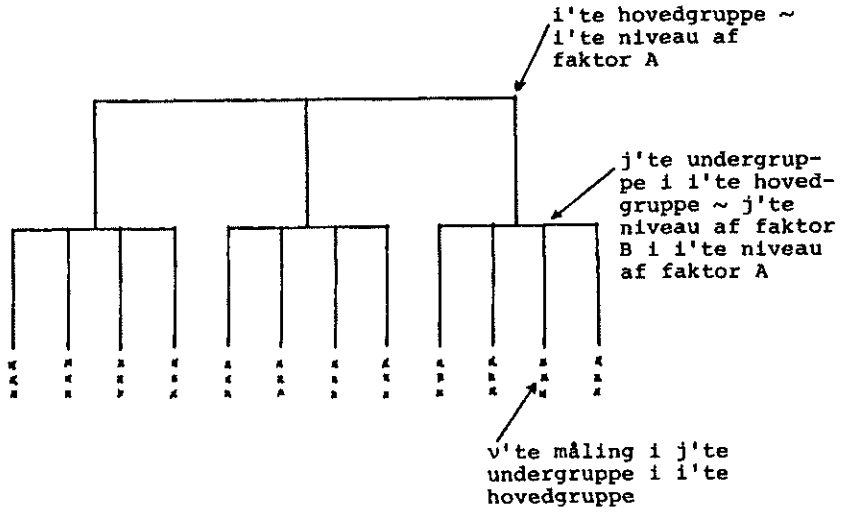
Overalt i resten af dette afsnit betragter vi observationer

$$\begin{array}{lrll} X_{ij\nu}, & i & = & 1,\ldots,k \quad \text{svarende til faktor A} \\ & j & = & 1,\ldots,m \quad \text{svarende til faktor B} \\ & \nu & = & 1,\ldots,n \quad \text{svarende til gentagelser} \end{array} .$$

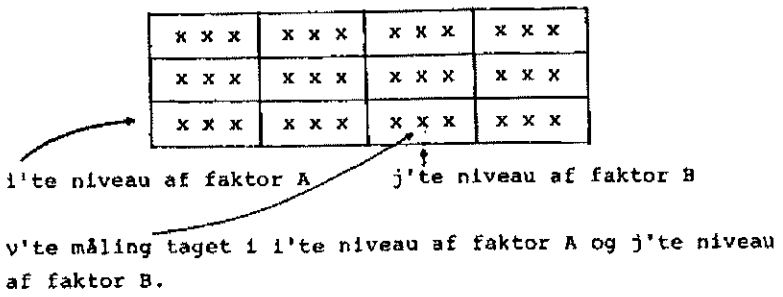
Vi taler om en **hierarkisk klassifikation**, hvis data "hænger" sammen som illustreret på figur 5.2.

Her er faktor B underordnet A, og vi skriver $A \supset B$. Hvis sammenhængen er som vist på figur 5.3.

Taler vi om en **krydsklassfikation**. Vi skriver $A \times B$.



Figur 5.2:



Figur 5.3:

I den **hierarkiske klassifikation** har vi følgende modeller

$$
\begin{array}{lrcl}
I: & X_{ij\nu} & = & \mu + a_i + b(a)_{j(i)} + z_{\nu(ij)} \\
II: & X_{ij\nu} & = & \mu + A_i + B(A)_{j(i)} + z_{\nu(ij)} \\
III: & X_{ij\nu} & = & \mu + a_i + B(a)_{j(i)} + z_{\nu(ij)}
\end{array}
$$

Vi har her valgt at følge den tidligere omtalte regel for indicering: de overordnede index er anført i en parentes.

Endvidere har vi angivet de enkelte parametre med samme bogstaver som de faktorer, de repræsenterer. Skal parameteren opfattes deterministisk, er valgt et lille bogstav; skal

den opfattes stokastisk er valgt et stort. **Optræder der flere bogstaver i en parameter, skal den opfattes stokastisk, blot ét af bogstaverne er stort.**

Bemærk, at når vi har valgt en betegnelse som $b(a)_{j(i)}$ angives ikke dermed et produkt. Der er blot tale om en parameter $\theta_{j(i)}$, som vi af mnemotekniske grunde kalder $b(a)_{j(i)}$.

Vi forudsætter, at

$$\sum_i a_i = 0$$

$$\sum_j b(a)_{j(i)} = 0 \qquad \forall i$$

$$\left.\begin{array}{l} Z_{\nu(ij)} \in \mathrm{N}(0,\sigma^2) \\ A_i \in \mathrm{N}(0,\sigma_A^2) \\ B(A)_{j(i)} \text{ (resp. } B(a)_{j(i)}) \in \mathrm{N}(0,\sigma_{B(A)}^2) \end{array}\right\} \text{ Stokastisk uafhængige}$$

Vi bemærker, at variansen svarende til $B(A)_{j(i)}$ og $B(a)_{j(i)}$ er indiceret $B(A)$, for at angive, at B er underordnet A.

Vi kalder I en **systematisk eller rent parametrisk model**, II en **stokastisk** eller en **varianskomponentmodel** og III en **blandet model**.

Man kan formelt også betragte blandede modeller af formen

$$X_{ij\nu} = \mu + A_i + b(a)_{j(i)} + Z_{\nu(ij)},$$

men dette forekommer ikke særlig rimeligt ("vekselvirkningen" er deterministisk og hovedeffekten stokastisk).

I den **krydsede klassifikation** har vi modellerne

$$\begin{array}{lrcl} I: & X_{ij\nu} & = & \mu + a_i + b_j + ab_{ij} + Z_{\nu(ij)} \\ II: & X_{ij\nu} & = & \mu + A_i + B_j + AB_{ij} + Z_{\nu(ij)} \\ III: & X_{ij\nu} & = & \mu + a_i + b_j + AB_{ij} + Z_{\nu(ij)} \\ IV: & X_{ij\nu} & = & \mu + a_i + B_j + aB_{ij} + Z_{\nu(ij)} \end{array}$$

Her er I en **systematisk** (eller rent parametrisk) model, II en **tilfældig** (eller en **varianskomponent**) model, og III og IV **blandede** modeller. Der findes naturligvis også en til IV analog model, hvor a_i er erstattet med A_i og B_j med b_j. Som i det hierarkiske tilfælde vil vi ikke betragte modeller, hvor vekselvirkningen er systematisk, og hovedvirkningen er tilfældig. I tilfældet med siloerne ville dette f.eks. svare til, at vi regner med, at vandindholdet i den i'te silo er "tilfældigt", f.eks. svarende til en tilfældig stikprøveudtagning blandt alle siloer i et givet område. I denne situation ville det virke barokt at påstå, at den afvigelse vi opnår fra den gennemsnitlige værdi ved at måle et

tilfældigt valgt sted i en tilfældigt valgt silo, naturligt beskrives ved en deterministisk parameterværdi.

Vi har følgende bånd og forudsætning er om fordelingerne

$$\sum_i a_i = 0$$

$$\sum_j b_j = 0$$

$$\sum_j ab_{ij} = 0 \qquad \forall_i$$

$$\sum_i ab_{ij} = 0 \qquad \forall_j$$

$$\left.\begin{array}{l} A_i \in \mathrm{N}(0, \sigma_A^2) \\ B_j \in \mathrm{N}(0, \sigma_B^2) \\ aB_{ij}, AB_{ij} \in \mathrm{N}(0, \sigma_{AB}^2) \\ Z_{p(ij)} \in \mathrm{N}(0, \sigma^2) \end{array}\right\} \text{stokastisk uafhængige}$$

Vi vil nu gå over til den statistiske analyse af de to typer af modeller. Begrundelserne for metoderne er analoge til dem, vi anvendte under den ensidede variansanalyse, så de forbigås og vi nøjes med at anføre resultaterne.

5.3.2 Analyse af hierarkisk klassificerede data

Grundlaget for de statistiske analyser kan findes i følgende variansanalysetabel med forventningsværdier over middelkvadratafvigelsessummer.

Vi har anvendt betegnelserne

$$\bar{X}_{ij\cdot} = \tfrac{1}{n}\textstyle\sum_\nu X_{ij\nu}$$

$$\bar{X}_{i\cdot\cdot} = \tfrac{1}{mn}\textstyle\sum_j \sum_\nu X_{ij\nu}$$

$$\bar{X} = \tfrac{1}{kmn}\textstyle\sum_i \sum_j \sum_\nu X_{ij\nu}.$$

Variation	SAK	Fr.gr.
Mellem hovedgrupper	$\mathrm{SAK}_A = nm\sum_i(\bar{X}_{i\cdot\cdot} - \bar{X})^2$	$f_A = k-1$
Mellem undergrupper inden for hovedgrupper	$\mathrm{SAK}_{B(A)} = n\sum_i\sum_j(\bar{X}_{ij\cdot} - \bar{X}_{i\cdot\cdot})^2$	$f_{B(A)} = k(m-1)$
Residual	$\mathrm{SAK}_{\mathrm{Res}} = \sum_i\sum_j\sum_\nu(X_{ij\nu} - \bar{X}_{ij\cdot})^2$	$f_{\mathrm{Res}} = km(n-1)$
Total	$\mathrm{SAK}_0 = \sum_i\sum_j\sum_\nu(X_{ij\nu} - \bar{X})^2$	$kmn-1$

	Systematisk model $X_{ij\nu} = \mu + a_i + b(a)_{j(i)} + Z_{\nu(ij)}$	Tilfældig model $X_{ij\nu} = \mu + A_i + B(A)_{j(i)} + Z_{\nu(ij)}$	Blandet model $X_{ij\nu} = \mu + A_i + B(a)_{j(i)} + Z_{\nu(ij)}$
$\mathrm{E}(\frac{\mathrm{SAK}_A}{f_A})$	$\sigma^2 + mn\frac{1}{k-1}\sum_i a_i^2$	$\sigma^2 + n\sigma^2_{B(A)} + mn\sigma^2_A$	$\sigma^2 + n\sigma^2_{B(A)} + mn\frac{1}{k-1}\sum a_i^2$
$\mathrm{E}(\frac{\mathrm{SAK}_{B(A)}}{f_{B(A)}})$	$\sigma^2 + \frac{n}{k(m-1)}\sum_{i,j} b(a)^2_{j(i)}$	$\sigma^2 + n\sigma^2_{B(A)}$	$\sigma^2 + n\sigma^2_{B(A)}$
$\mathrm{E}(\frac{\mathrm{SAK}_{\mathrm{Res}}}{f_{\mathrm{Res}}})$	σ^2	σ^2	σ^2

Hvis vi e.g. i den tilfældige model ønsker at teste hypotesen

$$H_0 : \sigma_A^2 = 0 \quad \text{mod} \quad H_1 : \sigma_A^2 > 0,$$

$$\{(x_{111},\ldots,x_{kmn}) | \frac{\mathrm{sak}_A/(k-1)}{\mathrm{sak}_{B(A)}/k(m-1)} > \mathrm{F}(k-1,k(m-1))_{1-\alpha}\}.$$

At

$$\frac{\mathrm{SAK}_A/(k-1)}{\mathrm{SAK}_{B(A)}/k(m-1)} \in \mathrm{F}(k-1,k(m-1)),$$

hvis $\sigma_A^2 = 0$, indses ved analoge betragtninger til de p. 223 gjorte. At de kritiske værdier er store værdier af brøken, indses let ved at betragte de forventede værdier.

Hvis vi i den systematiske model ønsker at teste den tilsvarende hypotese

$$H_0 : a_1 = \cdots = a_k = 0 \quad \text{mod} \quad H_1 : \exists i,j : a_i \neq a_j,$$

bliver det kritiske område

$$\{(x_{111},\ldots,x_{kmn}) | \frac{\mathrm{sak}_A/(k-1)}{\mathrm{sak}_{\mathrm{Res}}/km(n-1)} > \mathrm{F}(k-1,km(n-1))_{1-\alpha}\}.$$

Her bemærker vi en **forskel** mellem analysen af den systematiske model og den tilfældige.

Ved hjælp af tabellen over de forventede værdier konstruerer man nu let de kritiske områder for de øvrige hypoteser, man måtte ønske at teste.

Inden vi giver et eksempel, skal vi anføre nogle estimatorer:

$$\begin{aligned} \hat{\mu} &= \bar{X} \\ \hat{a}_i &= \bar{X}_{i\cdot\cdot} - \bar{X} \\ \hat{b}(a)_{j(i)} &= \bar{X}_{ij\cdot} - \bar{X}_{i\cdot\cdot} \end{aligned}$$

Estimatorerne for de forskellige varianser fås som summer og differenser af de forskellige kvadratafvigelsessummer. I den tilfældige model haves e.g., at

$$\hat{\sigma}_A^2 = \frac{1}{mn}(\mathrm{SAK}_A/f_A - \mathrm{SAK}_{B(A)}/f_{B(A)})$$

er et centralt skøn over σ_A^2.

Vi anfører nu et eksempel, hvor data er taget fra [15], her citeret fra [16].

EKSEMPEL 5.4. Ved en undersøgelse af 60 mm morterer har man testet 5 partier ammunition. Hvert parti blev testet 2 forskellige dage på grund af eksperimentets størrelse. I nedenstående tabel er der anført observeret rækkevidde (i yards) for granater af typen HE, M49A2 affyret med standardladning ved en elevation på 45%.

Det er klart, at vi her har en hierarkisk model. Hovedgrupperne er ammunitionspartierne, og undergrupperne er dagnumrene (grunden til, at vi ikke har en krydsklassifikation er, at dagene er forskellige. Hvis vi havde prøvet samtlige partier e.g. 24/10-52 og igen samtlige partier den 31/10-52, havde vi haft en krydsklassifikation).

Vi vælger modellen

$$X_{ij\nu} = \mu + a_i + B(a)_{j(i)} + Z_{\nu(ij)},$$

hvor

i	$=$	$1,\ldots,5$	(parti)
j	$=$	$1,2$	(dag inden for parti)
ν	$=$	$1,\ldots,19$	(gentagelse inden for parti og dag)

og

Parti	MA−1 − 53		MA−1 − 82		MA−1 − 363		MA−1 − 604		MA−613	
Dato	24/10 52	31/10 52	17/12 52	12/1 53	19−20/8 53	8 − 9/9 53	30/12 54	10/1 55	20/1 55	31/1 55
	1890	2057	1925	2112	1988	1932	1967	1980	2000	2110
	1863	2028	1902	2083	1876	1862	2021	2025	1769	1983
	1927	1964	2043	2096	1874	1863	2014	1983	1885	2098
	1830	1955	1957	2078	1914	1927	2019	1862	2004	2084
	1803	1976	1946	2031	1882	1907	2002	2041	1904	2015
	1951	1996	1940	2084	1774	1763	2128	2001	1865	1978
Række-	1995	2057	1916	2017	1872	1841	1949	1970	1927	2098
vidde	1967	1979	1967	2035	1822	1914	1904	2053	1972	2124
i	1934	2010	1958	1978	1891	1837	2029	1978	1886	2077
yards	1897	2037	1879	2045	1855	1911	1989	1940	2019	2036
	1869	2013	1995	2002	1809	1866	2052	1980	1884	2060
	1847	1990	1980	2078	1894	1797	2042	1921	1990	2141
	1882	2015	1934	2118	1870	1983	1835	2002	1884	2075
	1965	1975	1990	2017	1910	1873	1970	1969	1938	2074
	1954	1934	1906	2107	1970	1907	2038	1849	1950	2003
	1973	2074	1985	2005	1980	1923	2000	2030	1999	2077
	1870	2071	2000	2094	1885	1962	2041	2006	1987	2110
	1894	1993	1951	1993	1775	1859	2033	1999	1823	2061
	1927	1943	1979	2020	1871	1993	1996	2006	1951	2099

$$a_1 + \cdots + a_5 = 0$$

$$\left.\begin{array}{ll} B(a)_{j(i)} \in & \mathrm{N}(0, \sigma^2_{B(A)}) \\ Z_{\nu(ij)} \in & \mathrm{N}(0, \sigma^2) \end{array}\right\} \text{ stokastiske uafhængige.}$$

Grunden til, at vi har valgt en blandet model er, at den j'te dags afvigelse fra det pågældende partigennemsnit er sammensat af en række **tilfældige** bidrag, hvoraf det væsentligste formentlig er de meteorologiske forhold. Endvidere er vi næppe særskilt interesserede i at analysere forholdet de angivne dage, men nok snarere interesserede i at arbejde med en model,der er egnet til at forudsige,hvad man kan forvente ved en gentagelse af eksperimentet på helt andre dage. Det ville derfor ikke være rimeligt at anvende en rent systematisk model.

Det primære må nu være at undersøge, om alle a_i'erne kan antages at være 0, i.e. om det kan antages, at ammunitionspartierne er ens. Sekundært vil man være interesseret i at undersøge om $\sigma^2_{B(A)}$ kan antages at være 0. Hvis dette er tilfældet, indebærer det bl.a., at man ved fremtidige undersøgelser af samme art f.eks. ikke behøver at tage hensyn til, at forsøgene strækker sig over flere dage.

Vi samler beregningerne i følgende variansanalyseskema. (Vi skal senere vende tilbage til beregningsspørgsmålet).

Variation	SAK	f	S^2	$\mathrm{E}(S^2)$
Mellem partier	369254	4	92314	$\sigma^2 + 19\sigma^2_{B(A)} + 38 \cdot \frac{1}{4}\sum a_i^2$
Mellem dage inden for partier	369941	5	73988	$\sigma^2 + 19\sigma^2_{B(A)}$
Inden for dage	517065	180	2873	σ^2
Total	1256260	189		

Da

$$\frac{92314}{73988} = 1.25 \simeq \mathrm{F}(4,5)_{0.59},$$

og

$$\frac{73988}{2873} = 25.75 > \mathrm{F}(5,180)_{0.9995},$$

ser vi, at hypotesen alle $a_i = 0$ accepteres på alle niveauer mindre end 41%, hvorimod hypotesen $\sigma_t^2 = 0$ forkastes meget kraftigt. (At det er de anførte brøker, der er de relevante, ses let ved hjælp af søjlen over de forventede værdier af S^2-størrelserne).

Vor konklusion må derfor blive, at der ikke skønnes at være nogen (signifikant) forskel mellem ammunitionspartierne. Derimod er der en tydelig dag-effekt. Den estimerede varians for denne er

$$\hat{\sigma}^2_{B(A)} = \frac{1}{19}(73988 - 2873) = 3742.89 = 61.18^2.$$

Endelig har vi følgende estimat over forsøgsfejlens varians:

$$\hat{\sigma}^2 = 2873 = 53.60^2.$$

Et skøn over variansen på en enkelt måling er derfor lig

$$\hat{\sigma}^2 + \hat{\sigma}^2_{B(A)} = 6616 = 81.3^2.$$

♦

BEMÆRKNING 5.1. Hvis vi - noget søgt - havde anvendt en rent systematisk model, havde vi ved test af hypotesen om alle $b(a)_{j(i)} = 0$ fået teststørrelsen

$$\frac{73988}{2873} = 25.75 > \mathrm{F}(5,180)_{0.9995},$$

og ved test af hypotesen alle $a_i = 0$ teststørrelsen

$$\frac{92314}{2873} = 32.13 > \mathrm{F}(4, 180)_{0.9995}$$

I denne situation må vor konklusion derfor blive, at såvel $b(a)_{j(i)}$'er som a_i'er (NB!) må antages at være forskellige fra 0, altså tilsyneladende et resultat, der er i modstrid med det tidligere opnåede. Deri ligger dog ingen egentlig modstrid; thi modellerne er forskellige, og vi tolker afvigelser (variationer) forskelligt i de to situationer. ▼

5.3.3 Analyse af krydsklassificerede data

Variansanalyseskemaet er her det fra den almindelige tosidede variansanalyse (d.v.s. den systematiske model) velkendte. Foruden dette anfører vi også en oversigt over de forventede værdier af SAK_i/f_i. (Vi benævner som sædvanlig index i som rækkeindex og index j som søjleindex):

Variation	SAK	Frihedsgr.
Mellem rækker (faktor A)	$\mathrm{SAK}_A = mn\sum_i(\bar{X}_{i\cdot\cdot} - \bar{X})^2$	$k-1$
Mellem søjler (faktor B)	$\mathrm{SAK}_B = kn\sum_j(\bar{X}_{\cdot j\cdot} - \bar{X})^2$	$m-1$
Vekselvirkning (AB)	$\mathrm{SAK}_{AB} = n\sum_i\sum_j(\bar{X}_{ij\cdot} - \bar{X}_{i\cdot\cdot} - \bar{X}_{\cdot j\cdot} + \bar{X})^2$	$(k-1)(m-1)$
Residual	$\mathrm{SAK}_{\mathrm{Res}} = \sum_i\sum_j\sum_\nu(X_{ij\nu} - \bar{X}_{ij\cdot})^2$	$km(n-1)$
Total	$\sum_i\sum_j\sum_\nu(X_{ij\nu} - \bar{X})^2$	$kmn-1$

	Systematisk model $X_{ij\nu} = \mu + a_i + b_j + ab_{ij} + Z_{\nu(ij)}$	Tilfældig model $X_{ij\nu} = \mu + A_i + B_j + AB_{ij} + Z_{\nu(ij)}$
$\mathrm{E}(\frac{\mathrm{SAK}_A}{f_A})$	$\sigma^2 + mn\frac{1}{k-1}\sum_i a_i^2$	$\sigma^2 + n\sigma^2_{AB} + mn\sigma^2_A$
$\mathrm{E}(\frac{\mathrm{SAK}_B}{f_B})$	$\sigma^2 + kn\frac{1}{m-1}\sum_j b_j^2$	$\sigma^2 + n\sigma^2_{AB} + kn\sigma^2_B$
$\mathrm{E}(\frac{\mathrm{SAK}_{AB}}{f_{AB}})$	$\sigma^2 + n\frac{1}{(k-1)(m-1)}\sum_{i,j} ab_{ij}^2$	$\sigma^2 + n\sigma^2_{AB}$
$\mathrm{E}(\frac{\mathrm{SAK}_{\mathrm{Res}}}{f_{\mathrm{Res}}})$	σ^2	σ^2

	Blandet model $X_{ij\nu} = \mu + a_i + b_j + AB_{ij} + Z_{\nu(ij)}$	Blandet model $X_{ij\nu} = \mu + a_i + B_j + aB_{ij} + Z_{\nu(ij)}$
$\mathrm{E}(\frac{\mathrm{SAK}_A}{f_A})$	$\sigma^2 + n\sigma^2_{AB} + mn\frac{1}{k-1}\sum_i a_i^2$	$\sigma^2 + n\sigma^2_{AB} + mn\frac{1}{k-1}\sum_i a_i^1$
$\mathrm{E}(\frac{\mathrm{SAK}_B}{f_B})$	$\sigma^2 + n\sigma^2_{AB} + kn\frac{1}{m-1}\sum_j b_j^2$	$\sigma^2 + n\sigma^2_{AB} + kn\sigma^2_B$
$\mathrm{E}(\frac{\mathrm{SAK}_{AB}}{f_{AB}})$	$\sigma^2 + n\sigma^2_{AB}$	$\sigma^2 + n\sigma^2_{AB}$
$\mathrm{E}(\frac{\mathrm{SAK}_{\mathrm{Res}}}{f_{\mathrm{Res}}})$	σ^2	σ^2

Estimatorerne for de systematiske parametre er

$$\begin{aligned}\hat{\mu} &= \bar{X}\\ \hat{a}_i &= \bar{X}_{i\cdot\cdot} - \bar{X}\\ \hat{b}_j &= \bar{X}_{\cdot j\cdot} - \bar{X}\\ \hat{ab}_{ij} &= \bar{X}_{ij\cdot} - \bar{X}_{i\cdot\cdot} - \bar{X}_{\cdot j\cdot} + \bar{X}_{\cdot\cdot\cdot}\end{aligned}$$

Varianserne estimeres som vanligt ved passende summer og differenser mellem SAK_i/f_i'er. Teststørrelserne for de forskellige hypoteser er kvotienter mellem SAK_i/f_i'er. Af de forventede værdier ses umiddelbart, hvilke der kan komme på tale.

Vi illustrerer anvendelserne af disse skemaer ved hjælp af følgende

EKSEMPEL 5.5. (Data stammer fra reference [6]?bog). På en klædefabrik har man undersøgt 4 spolemaskiners produktion. Nærmere betegnet har man til 3 forskellige tidspunkter taget 4 tråde af fast længde og vejet disse. Resultaterne fremgår af følgende tabel

		Maskine (faktor B)			
		1	2	3	4
Tidspunkt (faktor A)	α	7.50	7.81	7.50	8.05
		7.70	7.94	7.77	7.53
		7.93	7.23	7.96	8.16
		7.52	7.94	7.83	7.76
	β	8.17	8.29	8.46	8.38
		8.09	8.54	8.33	8.47
		8.11	8.45	8.27	8.38
		7.96	8.43	8.24	8.60
	γ	8.01	8.01	8.37	8.16
		8.17	7.92	8.27	7.96
		8.05	8.27	8.07	8.08
		7.91	7.92	8.28	8.52

Man er åbenbart interesseret i at få beskrevet variationen mellem maskinerne. Afhængigt af, hvorledes tidspunkterne er valgte, er der 2 mulige modeller.

Hvis tidspunkterne er systematisk valgt (α f.eks. $\frac{1}{2}$time efter arbejdstids start, $\beta = \frac{1}{2}$ time før frokost og $\gamma = \frac{1}{2}$ time før fyraften), vil vi anvende modellen

$$I: \qquad X_{ij} = \mu + a_i + b_j + ab_{ij} + Z_{\nu(ij)} \qquad \begin{array}{l} i = 1,2,3 \\ j = 1,2,3,4 \\ v = 1,2,3,4, \end{array}$$

hvor μ angiver et niveau, a_i angiver tidseffekten, b_j maskineffekten og ab_{ij} vekselvirkningen.

Hvis tidspunkterne er tilfældigt valgte, bør vi anvende modellen

$$II: \qquad X_{ij\nu} = \mu + A_i + b_j + Ab_{ij} + Z_{\nu(ij)},$$

idet vi nu tolker tidseffekten og vekselvirkningen som stokastiske variable.

Variansanalyseskemaet bliver

Variation	SAK	Fr.gr.	S^2	$\mathrm{E}(S^2)$ I	$\mathrm{E}(S^2)$ II
Mellem tider (A)	2.6264	2	1.3132	$\sigma^2 + 8\sum_i a_i^2$	$\sigma^2 + 4\sigma_{AB}^2 + 16\sigma_A^2$
Mellem mask. (B)	0.3907	3	0.13023	$\sigma^2 + 4\sum_j b_j^2$	$\sigma^2 + 4\sigma_{AB}^2 + 4\sum_j b_j^2$
Vekselvirkn. (AB)	0.1918	6	0.03197	$\sigma^2 + \frac{2}{3}\sum_{i,j} ab_{ij}^2$	$\sigma^2 + 4\sigma_{AB}^2$
Residual (Res)	1.2768	36	0.03547	σ^2	σ^2
Total	4.4857	47			

Teststørrelsen for, om $ab_{ij} = 0$ i den parametriske model, respektive, om $\sigma_{AB}^2 = 0$ i den blandede model, er

$$\frac{0.03197}{0.03547} = 0.90 \simeq \mathrm{F}(6,36)_{0.50}.$$

Vi vil derfor acceptere, at $ab_{ij} = 0$ henholdsvis $\sigma_{AB}^2 = 0$ ved test på alle niveauer mindre end 50%. Vi antager derfor dette, og får som et forbedret skøn over σ^2

$$\hat{\sigma}^2 = \frac{0.1918 + 1.2768}{6 + 36} = 0.0350.$$

Vi ser nu, at teststørrelserne for test om $a_i = 0$ og $b_j = 0$ henholdsvis $\sigma_a^2 = 0$ og $\beta_j = 0$ bliver ens i de to modeller, nemlig

$$\frac{S^2_{\text{tid}}}{0.0350} = 37.52 \simeq \mathrm{F}(2,42)_{0.9995}$$
$$\frac{S^2_{\text{maskine}}}{0.0350} = 3.72 \simeq \mathrm{F}(3,42)_{0.98}.$$

Vi vil åbenbart forkaste hypotesen om, at a_i'erne = 0 respektive $\sigma_A^2 = 0$. Hypotesen om, at b_j'erne er lig 0, er mere tvivlsom.

Hvis tidspunkterne er tilfældigt valgte (d.v.s. vi regner med model II), bliver konklusionen, at det er tvivlsomt, om der er forskel på maskinerne. Der spores derimod en betydelig variation mellem målinger taget til forskellige tidspunkter. Variansen på disse variationer estimeres ved

$$\hat{\sigma}_A^2 = \frac{1}{16}(1.3132 - 0.0350) = 0.0799 = 0.2826^2.$$

Niveauet estimeres til

$$\hat{\mu} = 8.068,$$

og "forsøgsfejlens" varians til

$$\hat{\sigma}^2 = 0.0350 = 0.187^2,$$

der er estimeret med 42 frihedsgrader.

Er tidspunkternes systematisk valgte, må vi fremdeles konkludere, at det er tvivlsomt, om der er forskel på maskinerne. Derimod er der en tydelig tidseffekt, der udtrykkes ved estimaterne

$$\begin{aligned} \hat{a}_1 &= -0.310 \\ \hat{a}_2 &= 0.255 \\ \hat{a}_3 &= 0.055. \end{aligned}$$

Niveauet estimeres til

$$\hat{\mu} = 8.068,$$

og forsøgsfejlens varians (med 42 frihedsgrader) til

$$\hat{\sigma}^2 = 0.0350 = 0.187^2.$$

♦

Ved sammenligning af variansanalyseskemaerne for den hierarkiske model og krydsklassifikationsmodellen ses, at SAK for variationen mellem hovedgrupper i den hierarkiske model er den samme som SAK for variationen mellem rækker i krydsmodellen. Endvidere er residual-SAK'erne og de totale SAK'er ens. Heraf fås umiddelbart, at

$$\mathrm{SAK}_{B(A)} = \mathrm{SAK}_B + \mathrm{SAK}_{AB}$$

eller udtrykt mere direkte

$$n\sum_i\sum_j(\bar{X}_{ij\cdot} - \bar{X}_{i\cdot})^2 =$$
$$kn\sum_j(\bar{X}_{\cdot j\cdot} - \bar{X})^2 + n\sum_i\sum_j(\bar{X}_{ij\cdot} - \bar{X}_{i\cdot} - \bar{X}_{\cdot j\cdot} + \bar{X})^2$$

(Denne relation vises også let direkte).

Dette er væsentligt, fordi der findes let tilgængelige standardprogrammer til beregning af kvadratsummerne i en almindelig (krydsklassificeret) tosidet variansanalyse. Man kan derfor køre sin hierarkiske model som en krydsklassificeret og så få den korrekte variansanalysetabel ved blot at addere to kvadratsummer. Vi bemærker, at denne additionsregel også er gyldig for frihedsgraderne.

Regnes eksempel 5.4 som en krydset tosidet variansanalyse, fås følgende tabel

Variation	SAK		
Mellem rækker	369254		4
Mellem søjler	197000	369941	1
Vekselvirkning	172941		4
Residual	517065		180
Total	1256260		189

Det ses, at vi ved addition af SAK (mellem søjler) og SAK (vekselvirkning) netop får det oprindelige variansanalyseskema.

Et andet meget afgørende forhold er spørgsmålet med at bestemme den "fejlvarians", som effekterne skal måles relativt til.

I den krydsede model med **deterministiske** parametre skal vi bruge residualkvadratafvigelsessummen. Hvis vi kun har en observation pr. celle er denne lig 0. Under **forudsætning** af, at der **ikke** er nogen **vekselvirkning** (i.e. at alle $ab_{ij} = 0$) kan vi bruge vekselvirkningskvadratafvigelsessummen.

Har vi derimod en model med tilfældig vekselvirkning, kan vi altid bruge vekselvirkningskvadratafvigelsessummen. En sådan model kunne f.eks. svare til et forsøg som følgende.

EKSEMPEL 5.6. Ved en analyse af forskellige diæters indflydelse på personers vægttab for forskellige aldersgrupper har man opnået følgende måleresultater:

		B: Alder		
		Person på 20 år	$\cdots$	Person på 50 år
	1	$X_{111},\ldots,X_{115}$	$\cdots$	$X_{141},\ldots,X_{145}$
A : Diæt	$\vdots$	$\vdots$		$\vdots$
	5	$X_{511},\ldots,X_{515}$	$\cdots$	$X_{541},\ldots,X_{545}$

Ved bestemmelsen af vægttabet har man taget 5 uafhængige målinger af vægten.

Her har vi en model

$$X_{ij\nu} = \mu + a_i + b_j + AB_{ij} + Z_{\nu(ij)},$$

hvor

$$\sum a_i = \sum b_j = 0$$

$$\left.\begin{array}{ll} AB_{ij} \in & \mathrm{N}(0,\sigma^2_{AB}) \\ Z_{\nu(ij)} \in & \mathrm{N}(0,\sigma^2) \end{array}\right\} \text{ uafhængige.}$$

Vi regner begge hovedeffekterne deterministiske, den ene svarer til en bestemt diæt, den anden til en bestemt aldersgruppe. Derimod fortolker vi afvigelserne fra denne model som tilfældige. De kan e.g. skyldes det konkrete valg af forsøgspersoner etc.

Ved testningen af, om a_i'er eller b_j'er kan antages at være lig 0, er det klart, at vi ikke vil måle disse i forhold til σ^2, som er et udtryk for **veje**-usikkerheden; men derimod i forhold til den "**biologiske varians**" σ^2_{AB}, et resultat man også opnår ved anvendelse af den foran anførte teori. ♦

5.4 Variansanalysemodeller med 3 faktorer

I modeller, hvori der indgår mere end 2 faktorer, får vi mulighed for mere indviklede strukturer, hvor nogle faktorer er krydsede og andre er hierarkisk ordnede. Inden vi giver generelle formler, vil vi betragte nogle af de vigtigste muligheder med 3 faktorer A, B og C.

Vi kalder den almindelige observation

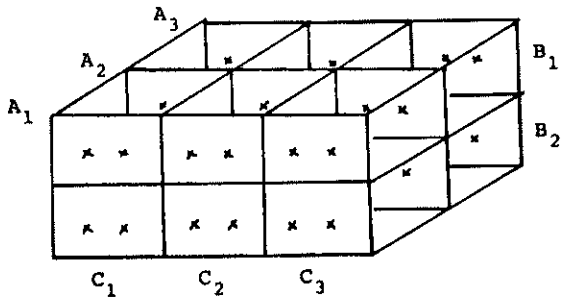
$$\begin{array}{lllll} X_{ijr\nu}, & i & = & l,\ldots,k & \text{angiver A-niveau} \\ & j & = & l,\ldots,m & \text{angiver B-niveau} \\ & r & = & l,\ldots,p & \text{angiver C-niveau} \\ & \nu & = & l,\ldots,n & \text{angiver gentagelsesnr.} \end{array}$$

Det grundlæggende variansanalyseskema for samtlige variansanalyser med de 3 faktorer er anført p. 245.

Vi betragter nu de enkelte strukturer.

Struktur 1 Alle faktorer krydsklassificerede, d.v.s. symbolsk $A \times B$, $A \times C$, $B \times C$.

Denne struktur kan grafisk anskueliggøres som i følgende skitse.



Med denne datastruktur har vi for det første den rent **systematiske** model

$$i) \quad X_{ijr\nu} = \mu + a_i + b_j + c_r + ab_{ij} + ac_{ir} + bc_{jr} + abc_{ijr} + Z_{\nu(ijr)}$$

og den rent **tilfældige** model

$$ii) \quad X_{ijr\nu} = \mu + A_i + B_j + C_r + AB_{ij} + AC_{ir} + BC_{jr} + ABC_{ijr} + Z_{\nu(ijr)}.$$

Endvidere forekommer der de **blandede** modeller, i.e. modeller hvor visse af komponenterne er konstanter og visse er stokastiske variable.

Parametrene og de stokastiske variable skal tilfredstille

$$\sum_i a_i = 0, \quad \sum_j b_j = 0, \quad \sum_r c_r = 0$$

$$\sum_i ab_{ij} = 0, \quad \sum_j ab_{ij} = 0; \quad \sum_i ac_{ir} = 0, \quad \sum_r ac_{ir} = 0;$$

$$\sum_j bc_{jr} = 0, \quad \sum_r bc_{jr} = 0.$$

$$\sum_i abc_{ijr} = 0, \quad \sum_j abc_{ijr} = 0, \quad \sum_r abc_{ijr} = 0.$$

$$A_i \in \mathrm{N}(0, \sigma_A^2), \quad B_j \in \mathrm{N}(0, \sigma_B^2), \quad C_r \in \mathrm{N}(0, \sigma_C^2)$$

$$AB_{ij} \in \mathrm{N}(0, \sigma_{AB}^2), \quad AC_{ir} \in \mathrm{N}(0, \sigma_{AC}^2), \quad BC_{jr} \in \mathrm{N}(0, \sigma_{BC}^2)$$

$$ABC_{ijr} \in \mathrm{N}(0, \sigma_{ABC}^2)$$

$$Z_{\nu(ijr)} \in \mathrm{N}(0, \sigma^2).$$

De stokastiske variable forudsættes endvidere af være uafhængige.

Vi har nu det nedenfor anførte variansanalyseskema.

Variation	SAK	f	E(SAK/f)
A	SAK_A	$k-1$	$\sigma^2 + (n\sigma_{ABC}^2 + mn\sigma_{AC}^2 + np\sigma_{AB}^2) + nmp\sigma_A^2$
B	SAK_B	$m-1$	$\sigma^2 + (n\sigma_{ABC}^2 + nk\sigma_{BC}^2 + np\sigma_{AB}^2) + nkp\sigma_B^2$
C	SAK_C	$p-1$	$\sigma^2 + (n\sigma_{ABC}^2 + nk\sigma_{BC}^2 + nm\sigma_{AC}^2) + nkm\sigma_C^2$
AB	SAK_{AB}	$(k-1)(m-1)$	$\sigma^2 + (n\sigma_{ABC}^2) + np\sigma_{AB}^2$
AC	SAK_{AC}	$(k-1)(p-1)$	$\sigma^2 + (n\sigma_{ABC}^2) + nm\sigma_{AC}^2$
BC	SAK_{BC}	$(m-1)(p-1)$	$\sigma^2 + (n\sigma_{ABC}^2) + nk\sigma_{BC}^2$
ABC	SAK_{ABC}	$(k-1)(m-1)(p-1)$	$\sigma^2 + n\sigma_{ABC}^2$
Res	$\mathrm{SAK}_{\mathrm{Res}}$	$kmp(n-1)$	σ^2
Tot	$\mathrm{SAK}_{\mathrm{Tot}}$	$kmpn-1$	

Her skal forventningsværdierne af SAK_i/f_i tolkes på en særlig måde. Hvis de pågæl-

Variations-årsag	Kvadratafvigelsessum	Frihedsgrader
A	$\mathrm{SAK}_A = mpn \sum_{i=1}^{k} (\bar{X}_{i\cdot\cdot} - \bar{X})^2$	$k-1$
B	$\mathrm{SAK}_B = kpn \sum_{j=1}^{m} (\bar{X}_{\cdot j\cdot} - \bar{X})^2$	$m-1$
C	$\mathrm{SAK}_C = kmn \sum_{r=1}^{p} (\bar{X}_{\cdot\cdot r} - \bar{X})^2$	$p-1$
AB	$\mathrm{SAK}_{AB} = pn \sum_{i=1}^{k} \sum_{j=1}^{m} (\bar{X}_{ij\cdot} - \bar{X}_{i\cdot\cdot} - \bar{X}_{\cdot j\cdot} + \bar{X})^2$	$(k-1)(m-1)$
AC	$\mathrm{SAK}_{AC} = mn \sum_{i=1}^{k} \sum_{r=1}^{p} (\bar{X}_{i\cdot r} - \bar{X}_{i\cdot\cdot} - \bar{X}_{\cdot\cdot r} + \bar{X})^2$	$(k-1)(p-1)$
BC	$\mathrm{SAK}_{BC} = kn \sum_{j=1}^{m} \sum_{r=1}^{p} (\bar{X}_{\cdot jr} - \bar{X}_{\cdot j\cdot} - \bar{X}_{\cdot\cdot r} + \bar{X})^2$	$(k-1)(p-1)$
ABC	$\mathrm{SAK}_{ABC} \sum_{i=1}^{k} \sum_{j=1}^{m} \sum_{r=1}^{p} (\bar{X}_{ijr} - \bar{X}_{ij\cdot} - \bar{X}_{i\cdot r} - \bar{X}_{\cdot jr} + \bar{X}_{i\cdot\cdot} + \bar{X}_{\cdot j\cdot} + \bar{X}_{\cdot\cdot r} - \bar{X})^2$	$(k-1)(m-1)(p-1)$
Residual	$\mathrm{SAK}_{\mathrm{Res}} = \sum_{i=1}^{k} \sum_{j=1}^{m} \sum_{r=1}^{p} \sum_{\nu=1}^{n} (X_{ijr\nu} - \bar{X}_{ijr})^2$	$kmp(n-1)$
Total	$\mathrm{SAK}_{\mathrm{Tot}} = \sum_{i=1}^{k} \sum_{j=1}^{m} \sum_{r=1}^{p} \sum_{\nu=1}^{h} (X_{ijr\nu} - \bar{X})$	$kmpn-1$

Tabel 5.1: Variationsanalyseskema for den fuldstændigt krydsede model med 3 faktorer.

dende effekter er systematiske, skal vi læse

$$\begin{aligned}
\sigma^2_{ABC} &= \frac{1}{(k-1)(m-1)(p-1)} \sum_i \sum_j \sum_r abc^2_{ijr} \\
\sigma^2_{BC} &= \frac{1}{(m-1)(p-1)} \sum_j \sum_r bc^2_{jr} \\
\sigma^2_{AC} &= \frac{1}{(k-1)(p-1)} \sum_i \sum_r ac^2_{ir} \\
\sigma^2_{AB} &= \frac{1}{(k-1)(m-1)} \sum_i \sum_j ab^2_{ij} \\
\sigma^2_C &= \frac{1}{p-1} \sum_r c^2_r \\
\sigma^2_B &= \frac{1}{m-1} \sum_j b^2_j \\
\sigma^2_A &= \frac{1}{k-1} \sum_i a^2_i .
\end{aligned}$$

De led i søjlen med forventningsværdierne, som er anbragt i parenteser, skal kun medtages, hvis de er udtryk for en tilfældig (stokastisk) effekt.

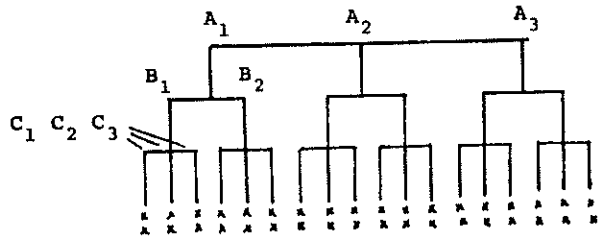
Hvis vi e.g. har en blandet model

$$X_{ijr\nu} = \mu + a_i + b_j + c_r + ab_{ij} + ac_{ir} + BC_{jr} + ABC_{ijr} + Z_{\nu(ijr)},$$

bliver eksempelvis

$$\mathrm{E}(\mathrm{SAK}_C/f_C) = \sigma^2 + n\sigma^2_{ABC} + nk\sigma^2_{BC} + nkm\frac{1}{p-1}\sum_r c^2_r .$$

Struktur 2 Den rent hierarkiske struktur, i.e. $A \supset B$, $B \supset C$, $A \supset C$. Strukturen er anskueliggjort i følgende skitse.



Den rent **systematiske** model bliver

$$X_{ijr\nu} = \mu + a_i + b(a)_{j(i)} + c(ab)_{r(ij)} + Z_{\nu(ijr)}.$$

Analogt skrives den **tilfældige** model

$$X_{ijr\nu} = \mu + A_i + B(A)_{j(i)} + C(AB)_{r(ij)} + Z_{\nu(ijr)},$$

og helt tilsvarende for de **blandede** modeller.

Variansanalyseskemaet bliver

Variation	SAK	f	E(SAK/f)
A	SAK_A	$k-1$	$\sigma^2 + (n\sigma^2_{C(AB)} + np\sigma^2_{B(A)}) + npm\sigma^2_A$
$B(A)$	$\mathrm{SAK}_B + \mathrm{SAK}_{AB}$	$k(m-1)$	$\sigma^2 + (n\sigma^2_{C(AB)}) + np\sigma^2_{B(A)}$
$C(AB)$	$\mathrm{SAK}_C + \mathrm{SAK}_{AC}$ $+\mathrm{SAK}_{BC} + \mathrm{SAK}_{ABC}$	$km(p-1)$	$\sigma^2 + n\sigma^2_{C(AB)}$
Res	$\mathrm{SAK}_{\mathrm{Res}}$	$kmp(n-1)$	σ^2
Tot	$\mathrm{SAK}_{\mathrm{Tot}}$	$kmpn-1$	

Hvis en effekt er systematisk, læses varianskomponenten efter nedenstående retningslinier

$$\begin{aligned}
\sigma^2_{C(AB)} &= \tfrac{1}{km(p-1)} \sum_i \sum_j \sum_r c(ab)^2_{r(ij)} \\
\sigma^2_{B(A)} &= \tfrac{1}{k(m-1)} \sum_i \sum_j b(a)^2_{j(i)} \\
\sigma^2_A &= \tfrac{1}{k-1} \sum_i a_i^2.
\end{aligned}$$

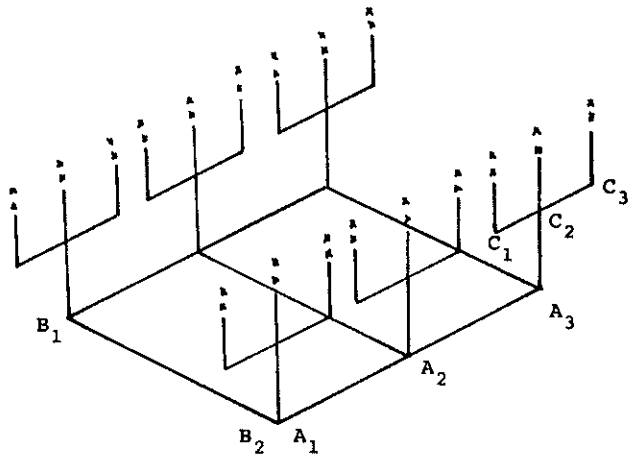
Angående parenteserne henvises til kommentaren p. 246.

NB! I de tilfældige modeller gælder de sædvanlige (i.e. de p. 242 nævnte) forudsætninger om de stokastiske komponenter. Der er imidlertid ændrede bånd på de systematiske, nemlig blot

$$\sum_r c(ab)_{r(ij)} = 0; \quad \sum_j b(a)_{j(i)} = 0; \quad \sum_i a_i = 0.$$

d.v.s. der er kun summer over index, der **ikke** står i en parantes, der er 0.

Struktur 3 De to faktorer krydsede og den tredie underordnet disse, d.v.s. $A \times B, A \supset C, B \supset C$. Strukturen er anskueliggjort grafisk nedenfor.



Den **systematiske** model er

$$X_{ijr\nu} = \mu + a_i + b_j + ab_{ij} + c(ab)_{r(ij)} + Z_{\nu(ijr)},$$

og den **tilfældige**

$$X_{ijr\nu} = \mu + A_i + B_j + AB_{ij} + C(AB)_{r(ij)} + Z_{\nu(ijr)}.$$

De **blandede** modeller dannes på vanlig måde. De ændrede bånd på parametrene er

$$\sum_r c(ab)_{r(ij)} = 0$$

$$\sum_i ab_{ij} = 0, \quad \sum_j ab_{ij} = 0$$

$$\sum_i a_i = 0, \quad \sum_j b_j = 0.$$

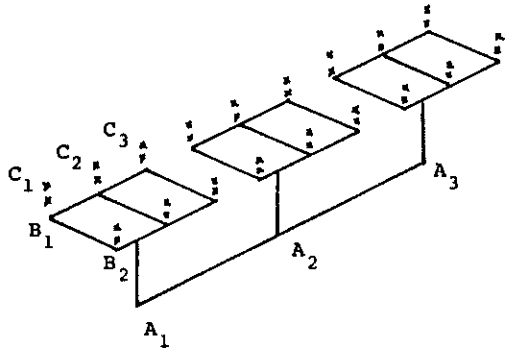
Variansanalyseskemaet er anført p. 248. Den deterministiske tolkning af varianskomponenterne er

$$\begin{aligned}
\sigma^2_{C(AB)} &= \tfrac{1}{km(p-1)} \sum_i \sum_j \sum_r c(ab)^2_{r(ij)} \\
\sigma^2_{AB} &= \tfrac{1}{(k-1)(m-1)} \sum_i \sum_j ab^2_{ij} \\
\sigma^2_B &= \tfrac{1}{m-1} \sum_j b_j^2 \\
\sigma^2_A &= \tfrac{1}{k-1} \sum_i a_i^2.
\end{aligned}$$

Variation	SAK	f	E(SAK/f)
A	SAK_A	$k-1$	$\sigma^2+(n\sigma^2_{C(AB)}+np\sigma^2_{AB})+npm\sigma^2_A$
B	SAK_B	$m-1$	$\sigma^2+(n\sigma^2_{C(AB)}+np\sigma^2_{AB})npk\sigma^2_B$
AB	SAK_{AB}	$(k-1)(m-1)$	$\sigma^2+(n\sigma^2_{C(AB)})+np\sigma^2_{AB}$
$C(AB)$	$\text{SAK}_C+\text{SAK}_{AC}$ $+\text{SAK}_{BC}+\text{SAK}_{ABC}$	$km(p-1)$	$\sigma^2+n\sigma^2_{C(AB)}$
Res	SAK_{Res}	$kmp(n-1)$	σ^2
Tot	SAK_{Tot}	$kmpn-1$	

Angående parenteserne gælder de sædvanlige bemærkninger (p. 246).

Struktur 4 To faktorer krydsklassificerede og begge underordnet en tredie faktor. Symbolsk $A \supset B$, $A \supset C$, $B \times C$. Strukturen er anskueliggjort grafisk nedenfor.



Vi har den **systematiske** model

$$X_{ijr\nu} = \mu + a_i + b(a)_{j(i)} + c(a)_{r(i)} + bc(a)_{jr(i)} + Z_{\nu(ijr)}.$$

Den **tilfældige** model er

$$X_{ijr\nu} = \mu + A_i + B(A)_{j(i)} + C(A)_{r(i)} + BC(A)_{jr(i)} + Z_{\nu(ijr)}.$$

Båndene på parametrene i den systematiske (og de **blandede**) modeller er

$$\sum_j bc(a)_{jr(i)} = 0, \quad \sum_r bc(a)_{jr(i)} = 0$$

$$\sum_r c(a)_{r(i)} = 0, \quad \sum_j b(a)_{j(i)} = 0$$

$$\sum_i a_i = 0.$$

Variansanalyseskemaet bliver

Variation	SAK	f	E(SAK/f)
A	SAK_A	$k-1$	$\sigma^2+(n\sigma^2_{BC(A)}+mn\sigma^2_{C(A)}$ $+pn\sigma^2_{B(A)})+mpn\sigma^2_A$
$B(A)$	$\mathrm{SAK}_B+\mathrm{SAK}_{AB}$	$k(m-1)$	$\sigma^2+(n\sigma^2_{BC(A)})+pn\sigma^2_{B(A)}$
$C(A)$	$\mathrm{SAK}_C+\mathrm{SAK}_{AC}$	$k(p-1)$	$\sigma^2+(n\sigma^2_{BC(A)})+mn\sigma^2_{C(A)}$
$BC(A)$	$\mathrm{SAK}_{BC}+\mathrm{SAK}_{ABC}$	$k(m-1)(p-1)$	$\sigma^2+n\sigma^2_{BC(A)}$
Res	$\mathrm{SAK}_{\mathrm{Res}}$	$kmp(n-1)$	σ^2
Total	$\mathrm{SAK}_{\mathrm{Tot}}$	$kmpn-1$	

Varianskomponenterne skal, såfremt de tilsvarende effekter er systematiske, læses som

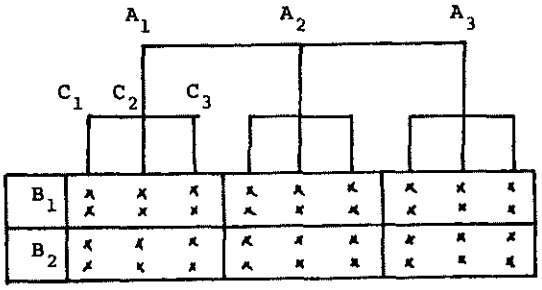
$$\begin{aligned}
\sigma^2_{BC(A)} &= \tfrac{1}{k(m-1)(p-1)}\sum_i\sum_j\sum_r bc(a)^2_{jr(i)}\\
\sigma^2_{C(A)} &= \tfrac{1}{k(p-1)}\sum_i\sum_r c(a)^2_{r(i)}\\
\sigma^2_{B(A)} &= \tfrac{1}{k(m-1)}\sum_i\sum_j b(a)^2_{j(i)}\\
\sigma^2_A &= \tfrac{1}{k-1}\sum_i a_i^2.
\end{aligned}$$

Vedrørende parenteserne henvises fremdeles til bemærkningen p. 246

Struktur 5 De to faktorer krydsede og den tredie underordnet den ene og krydset med den anden, d.v.s.

$$A\times B,\quad B\times C,\quad A\supset C.$$

Strukturen er søgt anskueliggjort i nedenstående skitse.



Den systematiske model er

$$X_{ijr\nu}=\mu+a_i+b_j+ab_{ij}+c(a)_{r(i)}+bc(a)_{jr(i)}+Z_{\nu(ijr)},$$

og den tilfældige

$$X_{ijr\nu} = \mu + A_i + B_j + AB_{ij} + C(A)_{r(i)} BC(A)_{jr(i)} + Z_{\nu(ijr)}.$$

parameterbåndene er

$$\sum_i a_i = \sum_j b_j = \sum_i ab_{ij} = \sum_j ab_{ij} = 0.$$
$$\sum_r c(a)_{r(i)} = \sum_r bc(a)_{jr(i)} = \sum_j bc(a)_{jr(i)} = 0.$$

Variansanalyseskemaet bliver

Variation	SAK	f	$\mathrm{E}(S^2)$
A	SAK_A	$k-1$	$\sigma^2 + (n\sigma^2_{BC(A)} + pn\sigma^2_{AB} + mn\sigma^2_{C(A)}) + mpn\sigma^2_A$
B	SAK_B	$m-1$	$\sigma^2 + (n\sigma^2_{BC(A)} + pn\sigma^2_{AB}) + kpn\sigma^2_B$
AB	SAK_{AB}	$(k-1)(m-1)$	$\sigma^2 + (n\sigma^2_{BC(A)}) + pn\sigma^2_{AB}$
$C(A)$	$\mathrm{SAK}_C + \mathrm{SAK}_{AC}$	$k(p-1)$	$\sigma^2 + (n\sigma^2_{BC(A)}) + mn\sigma^2_{C(A)}$
$BC(A)$	$\mathrm{SAK}_{BC} + \mathrm{SAK}_{ABC}$	$k(m-1)(p-1)$	$\sigma^2 + n\sigma^2_{BC(A)}$
Res	$\mathrm{SAK}_{\mathrm{Res}}$	$kmp(n-1)$	σ^2
Tot	$\mathrm{SAK}_{\mathrm{Tot}}$	$kmpn-1$	

Den deterministiske parameterfortolkning er

$$\begin{aligned}
\sigma^2_A &= \tfrac{1}{k-1}\sum_i a_i^2 \\
\sigma^2_B &= \tfrac{1}{m-1}\sum_j b_j^2 \\
\sigma^2_{AB} &= \tfrac{1}{(k-1)(m-1)}\sum_{i,j} ab_{ij}^2 \\
\sigma^2_{C(A)} &= \tfrac{1}{k(p-1)}\sum_{i,r} c(a)^2_{r(i)} \\
\sigma^2_{BC(A)} &= \tfrac{1}{k(m-1)(p-1)}\sum_{i,j,r} bc(a)^2_{jr(i)}
\end{aligned}$$

med den sædvanlige (p. 246) bemærkning om paranteserne.

5.5 Variansanalyser med flere faktorer

Vi indleder med et afsnit om

5.5.1 Estimation af parametre og beregning af kvadratafvigelsessummer

Vi har set, at variansanalysemodellerne med to og tre faktorer bygger på de fuldstændigt krydsede modeller. Helt analogt bygger de forskellige modeller med k faktorer på analysen af den fuldstændigt krydsede model med k faktorer.

Vi skal nu give en oversigt over fremgangsmåden. Faktorerne benævnes

$$F_1, \ldots, F_k,$$

men for at undgå en ekstra indicering vil vi ofte skrive

$$A, B, \ldots, G, H.$$

Antallet af niveauer for F_i sættes lig n_i. Svarende til disse niveauer forudsættes, at der foreligger observationer

$$\begin{array}{lll} X_{i_1 \ldots i_k \nu}, & i_1 = 1, \ldots, n_1 & (\sim \text{ faktor } F_1 = A) \\ & \vdots & \\ & i_k = 1, \ldots, n_k & (\sim \text{ faktor } F_k = H) \\ & \nu = 1, \ldots, n, & \end{array}$$

hvor ν angiver et gentagelsesindex.

I den deterministiske eller rent systematiske model antages, at middelværdierne i cellerne er spaltet som følger

$$\begin{aligned} \mathrm{E}(X_{i_1 \ldots i_k \nu}) &= \mu_{i_1 \ldots i_k} \\ &= \mu + a_{i_1} + \cdots + h_{i_k} \\ &\quad + ab_{i_1 i_2} + \cdots + gh_{i_{k-1} i_k} \\ &\quad \vdots \\ &\quad + a - - - g_{i_1 - - - i_{k-1}} + \cdots + b - - - h_{i_2 - - - i_k} \\ &\quad + a - - - h_{i_1 - - - i_k}. \end{aligned}$$

Her har vi også anvendt skrivemåden $a, b, \ldots, g, h$ i stedet for $f_1, f_2, \ldots, f_{k-1}, f_k$.

For en **vilkårlig af effekterne** og **af vekselvirkningerne** gælder, at **summen over et vilkårligt index er 0**.

Estimatorerne for parametrene bestemmes på sædvanlig måde. Man finder følgende

skøn, hvor vi samtidig har angivet den eller de faktorer, hvis effekter vi estimerer

$$
\begin{aligned}
A \quad &: \quad \hat{a}_{i_1} \quad = \quad \bar{X}_{i_1\ldots} - \bar{X} \\
&\vdots \\
H \quad &: \quad \hat{h}_{i_k} \quad = \quad \bar{X}_{\ldots i_k} - \bar{X} \\
AB \quad &: \quad \hat{ab}_{i_1 i_2} \quad = \quad \bar{X}_{i_1 i_2} - \bar{X}_{i_1\ldots} - \bar{X}_{.i_2\ldots} + \bar{X} \\
&\vdots \\
GH \quad &: \quad \hat{gh}_{i_{k-1} i_k} \quad = \quad \bar{X}_{\ldots i_{k-1} i_k} - \bar{X}_{\ldots i_{k-1}.} - \bar{X}_{\ldots i_k} + \bar{X} \\
ABC \quad &: \quad \widehat{abc}_{i_1 i_2 i_3} \\
&\qquad = \bar{X}_{i_1 i_2 i_3\ldots} - \bar{X}_{i_1 i_2\ldots} - \bar{X}_{i_1 . i_3\ldots} - \bar{X}_{.i_2 i_3\ldots} \\
&\qquad\quad + \bar{X}_{i_1\ldots} + \bar{X}_{.i_2\ldots} + \bar{X}_{..i_3\ldots} - \bar{X} \\
&\vdots \\
AB---G \quad &: \quad \widehat{ab---}\, g_{i_1---i_{k-1}} \\
= \quad & \bar{X}_{i_1---i_{k-1}.} - \bar{X}_{i_1---i_{k-2}..} - \bar{X}_{i_1---i_{k-3}.i_{k-1}.} - \cdots \\
& + (-1)^{k-3}\left[\bar{X}_{i_1 i_2\ldots} + \cdots + \bar{X}_{\ldots i_{k-2} i_{k-1}.}\right] \\
& + (-1)^{k-2}\left[\bar{X}_{i_1\ldots} + \cdots + \bar{X}_{\ldots i_{k-1}.}\right] + (-1)^{k-1}\bar{X} \\
&\vdots \\
AB---H \quad &: \quad \widehat{ab---}\, h_{i_1---i_k} \\
= \quad & \bar{X}_{i_1---i_i} - \bar{X}_{i_1---i_{k-1}.} - \cdots + (-1)^k \bar{X}.
\end{aligned}
$$

Et punktum på en indeksplads for X'ernes vedkommende betyder, at vi har summeret over det pågældende index. For at undgå forvekslinger med punktummer i en opremsning som a–h , er disse punktummer her erstattet med -.

Skønnet fremgår altså som gennemsnittet over de indices, der svarer til faktorer, der ikke indgår i effekten,
- alle de gennemsnit, der fremkommer ved at tage gennemsnit over yderligere et index,
+ alle de gennemsnit, der fremkommer ved at tage gennemsnit over yderligere 2 indices (i forhold til det første gennemsnit),
-
.
.
.
$+ (-1)^m \times$ totale gennemsnit.

For en vilkårlig faktorkombination $F_p \ldots F_r$ defineres nu den tilsvarende vekselvirkningskvadratafvigelsessum

$$\mathrm{SAK}_{F_p \ldots F_r} = \sum (f_p \widehat{---} f_r)^2_{i_p---i_r},$$

hvor summen udstrækkes over **samtlige k** indices $i_1, \ldots, i_k$.

Af hensyn til eventuelle beregninger kan det bemærkes, at kvadratsummerne kan skri-

ves som summer og differenser af kvadraterne på de led, der indgår i de enkelte vekselvirkningsestimater.

Eksempelvis er

$$\begin{aligned}\mathrm{SAK}_{F_1F_2F_3} &= \sum(\widehat{f_1f_2f_3})^2_{i_1i_2i_3} \\ &= \sum \bar{X}^2_{i_1i_2i_3\ldots} - \sum \bar{X}^2_{i_1i_2\ldots} - \sum \bar{X}^2_{i_1i_3\ldots} \\ &\quad - \sum \bar{X}^2_{i_2i_3\ldots} + \sum \bar{X}^2_{i_1\ldots} + \sum \hat{X}^2_{.i_2\ldots} \\ &\quad + \sum \bar{X}^2_{..i_3\ldots} - \sum \bar{X}^2.\end{aligned}$$

Her må det stadig erindres, at der summeres over samtlige indices. E.g. er

$$\sum \bar{X}^2 = n_1 \ldots n_k n \bar{X}^2.$$

Ved beregninger bør man af numeriske årsager heller ikke beregne de enkelte kvadratsummer direkte, men beregne kvadrater på summer og så foretage de nødvendige divisioner bagefter. E.g. er

$$\sum \bar{X}^2_{i_1i_2\ldots} = \frac{1}{n_3 \cdot n_4 \ldots n_k n} \sum_{i_1,i_2} \left(\sum_{i_3,\ldots,i_k,\nu} X_{i_1i_2i_3\cdots i_k\nu} \right)^2$$

Antallet af frihedsgrader for kvadratsummerne er

$$\mathrm{DF}(\mathrm{SAK}_{F_p---F_r}) = (n_p - 1)\cdots(n_r - 1).$$

Vi samler disse resultater i det grundlæggende variansanalyseskema for den fuldstændigt balancerede k-faktor variansanalysemodel. Skemaet er anført nedenfor.

Variationsårsag	SAK	Frihedsgrader
F_1	SAK_{F_1}	$n_1 - 1$
$\vdots$	$\vdots$	$\vdots$
F_k	SAK_{F_k}	$n_k - 1$
$F_1\ F_2$	$\mathrm{SAK}_{F_1\,F_2}$	$(n_1 - 1)(n_2 - 1)$
$\vdots$	$\vdots$	$\vdots$
$F_1 \ldots F_k$	$\mathrm{SAK}_{F_1\ldots F_k}$	$(n_1 - 1)\ldots(n_k - 1)$
Residual	$\mathrm{SAK}_{\mathrm{Res}}$	$n_1 \ldots n_k(n-1)$
Total	$\mathrm{SAK}_{\mathrm{Tot}}$	$n_1 \ldots n_k(n-1)$

Grundlæggende variansanalyseskema for balanceret k-faktorforsøg.

Forekommer der en hierarkisk ordning af visse faktorer eller en blanding af krydsning og hierarkisk ordning af faktorerne i modellen, kan såvel estimatorer som kvadratafvigelsessummer beregnes relativt enkelt ud fra den fuldstændigt krydsede model.

Lad os e.g. antage, at faktorer, der for simpelheds skyld benævnes $B_1, \ldots, B_r$ er underordnede faktorer $C_1, \ldots, C_s$.

Da finder vi for effekter svarende til $B_1 \ldots B_r(C_1 \cdots C_s)$

$$\begin{aligned}
&\{b_1 \cdots \widehat{b_r(c_1} \cdots c_s)\}_{i_1 \cdots i_r(j_1 \cdots j_s)} \\
&\quad = \{\widehat{b_1 \cdots b_r}\}_{i_1 \cdots i_r} + \{b_1 \widehat{\cdots b_r c_1}\}_{i_1 \cdots i_r j_1} + \cdots + \\
&\qquad + \{b_1 \cdots \widehat{b_r c_2} \cdots c_s\}_{i_1 \cdots i_r j_2 \cdots j_s} + \{b_1 \cdots \widehat{b_r c_1} \cdots c_s\}_{i_1 \cdots i_r j_1 \cdots j_s} \\
&\quad = \sum_{\{\ell_1, \ldots, \ell_t\} \subseteq \{1, \ldots, s\}} \{b_1 \cdots \widehat{b_r c_{\ell_1}} \cdots c_{\ell_t}\}_{i_1 \cdots i_r j_{\ell_1} \cdots j_{\ell_t}}.
\end{aligned}$$

Formlen synes meget kompliceret; men dens indhold er blot, at man beregner estimater i den fuldstændigt krydsede model, og derefter finder man estimater (i den hierarkiske model) for effekter svarende til
$B_1 \ldots B_r(C_1 \cdots C_s)$ ved at addere estimater for alle de effekter (i den krydsede model), der indeholder samtlige $B_1, \ldots, B_r$ og fra 0 til alle $C_1, \cdots, C_s$.

Vi bemærker i øvrigt, at de parameterbånd, der kommer på tale her, er, at summer over et vilkårligt index, der **ikke** står inde i en parentes, giver 0. Det ses umiddelbart, at også estimatorerne tilfredsstiller dette krav.

Skal man beregne kvadratafvigelsessummen svarende til ovennævnte effekt, sker det efter fuldstændigt samme regel, i.e.

$$\begin{aligned}
\mathrm{SAK}_{B_1 \cdots B_r(C_1 \cdots C_s)} &= \mathrm{SAK}_{B_1 \cdots B_r} + \mathrm{SAK}_{B_1 \cdots B_r C_1} + \cdots + \\
&\quad \mathrm{SAK}_{B_1 \cdots B_r C_2 \cdots C_s} + \mathrm{SAK}_{B_1 \cdots B_r C_1 \cdots C_s} \\
&= \sum_{\{\ell_1 \cdots \ell_t\} \subseteq \{1, \ldots, s\}} \mathrm{SAK}_{B_1 \cdots B_r C_{\ell_1} \cdots C_{\ell_t}}.
\end{aligned}$$

Antallet af frihedsgrader bliver summen af frihedsgraderne. Sættes antallet af niveauer for C_ν lig n_ν og antallet af niveauer for B_ν lig m, fås

$$\mathrm{DF}(\mathrm{SAK}_{B_1 \cdots B_r(C_1 \cdots C_s)}) = n_1 \cdots n_s(m_1 - 1) \cdots (m_r - 1).$$

5.5.2 Beregning af forventede værdier af middelkvadratafvigelsessummer

Ved testningen i en variansanalyse er det som sagt meget væsentligt at kunne beregne forventede værdier af middelkvadratafvigelsessummer.

Vi betragter en variansanalyse med faktorer

$$\ldots F_i \ldots; A; D_1, \ldots, D_t; B_1, \ldots, B_r; C_1, \ldots, C_s; R,$$

hvor R svarer til gentagelse. Antal niveauer for $F_i = n_i$.

Hvis der er tale om varianskomponentmodeller eller blandinger mellem varianskomponentmodeller og rent systematiske modeller, gør vi de sædvanlige forudsætninger, nemlig at alle involverede stokastiske variable er

1)	stokastisk uafhængige	3)	middelværdi 0 og med
2)	normalt fordelte med	4)	en varians, der evt. er 0.

Endvidere betragter vi i første omgang kun modeller, hvor det er således, at hvis en faktor er stokastisk, da er alle vekselvirkninger, hvor denne indgår, også stokastiske.

Der beregnes først et skema

Effekt	Faktor $\cdots\cdots F_i \cdots\cdots$	R
$\vdots$	$\vdots$	$\vdots$
A	$\delta_{F_i A}$	n
$\vdots$	$\vdots$	$\vdots$
$D_1 \ldots D_t$	$\delta_{F_i D_1 \ldots D_t}$	n
$\vdots$	$\vdots$	$\vdots$
$B_1 \ldots B_r(C_1 \ldots C_s)$	$\delta_{F_i B_1 \ldots B_r}(C_1 \ldots C_s)$	n
$\vdots$	$\vdots$	$\vdots$
$R(\cdots A \cdots C_s)$	$\cdots\cdots 1 \cdots\cdots$	1

Her er

$$\delta_{F_iA} = \begin{cases} n_i, & \text{hvis } F_i \neq A \\ 1, & \text{hvis } F_i = A \text{ og } F_i \text{ stokastisk} \\ 0, & \text{hvis } F_i = A \text{ og } F_i \text{ deterministisk} \end{cases}$$

$$\delta_{F_iD_1\cdots D_t} = \begin{cases} n_i, & \text{hvis } F_i \neq D_j \forall \quad j = 1, \ldots, t \\ 1, & \text{hvis } \exists j : F_i = D_j, \text{ og blot en af} \\ & D\text{'erne er stokastisk} \\ 0, & \text{hvis } \exists j : F_i = D_j, \text{ og alle } D\text{'er} \\ & \text{er deterministisk} \end{cases}$$

$$\delta_{F_iB_1\cdots B_r(C_1\cdots C_s)} = \begin{cases} n_i, & \text{hvis } \forall j : F_i \neq B_j \wedge \forall k : F_i \neq C_k \\ 1, & \text{hvis } \exists k : F_i = C_k \\ 1, & \text{hvis } \exists j : F_i = B_j \text{ og blot en af} \\ & B\text{'erne eller } C\text{'erne er stokasti-} \\ & \text{ske} \\ 0, & \text{hvis } \exists j : F_i = B_j \text{ og alle } B\text{'er og} \\ & C\text{'er er deterministiske.} \end{cases}$$

Ved den praktiske udfyldning af skemaet går man mest hensigtsmæssigt frem som følger. Vi betragter rækken svarende til en bestemt effekt.

1. Vi udsøger nu de faktorer, som ikke indgår i effekten. I de respektive søjler skrives antallet af niveauer for den tilsvarende faktor.
2. Dernæst udsøges de faktorer, som indgår i en **parentes** i effekten. I de respektive søjler skrives 1.
3. På de resterende pladser skrives 1, hvis effekten er stokastisk, og 0, hvis den er deterministisk.

Den forventede værdi af middelkvadratafvigelsessummen svarende til en bestemt effekt er af formen

$$\begin{aligned} \mathrm{E}(\mathrm{SAK}/f) &= \sigma^2 + \cdots + \alpha_1\varphi_A^2 + \cdots + \alpha_2\varphi_{D_1\cdots D_t}^2 \\ &\quad + \cdots + \alpha_3\varphi_{B_1\cdots B_r(C_1\cdots C_s)}^2 \end{aligned}$$

Her er α_i konstanter, der bestemmes i det følgende, og

$$\varphi_{\text{effekt}}^2 = \begin{cases} \sigma_{\text{effekt}}^2, & \text{hvis effekten er stokastisk} \\ \frac{1}{f_{\text{effekt}}} \sum(\text{parameter}_{\text{effekt}})^2, & \text{hvis effekten er} \\ & \text{deterministisk.} \end{cases}$$

α_i'erne fremkommer nu ved

i) at fjerne alle de søjler, der svarer til faktorer, der indgår i effekten,

ii) opsøge alle de rækker, der svarer til effekter, som indeholder **alle** de faktorer, der indgår i den bestemte effekt, vi undersøger.

iii) I hver af de betragtede rækker danner vi nu produktet af de led, der efter sletningen i i) er tilbage.

De derved fremkomne størrelser er koefficienterne α_i.

Efter således at have bestemt alle forventningsværdier, danner man F-tests for de relevante hypoteser på vanlig måde ved at beregne kvotienter mellem SAK/f'er.

Ofte vil det dog ikke være muligt at danne en teststørrelse umiddelbart, idet der muligvis ikke findes nogen SAK/f, der har en forventningsværdi, som direkte muliggør en testning af den interessante parameter.

Man vil da danne en linearkombination af flere SAK/f'er, som har den ønskede forventningsværdi, f.eks.

$$\mathrm{MS} = a_1\mathrm{MS}_1 + \cdots + a_k\mathrm{MS}_k,$$

hvor $\mathrm{MS}_i = \mathrm{SAK}_i/f_i$. Hvis $\mathrm{E}(\mathrm{MS}_i) = \sigma_i^2$, fås

$$\mathrm{E}(\mathrm{MS}) = a_1\sigma_1^2 + \cdots + a_k\sigma_k^2$$

og

$$\begin{aligned}\mathrm{V}(\mathrm{MS}) &= a_1^2\,\mathrm{V}(\mathrm{MS}_1) + \cdots + a_k^2\,\mathrm{V}(\mathrm{MS}_k)\\ &= 2[a_1^2\frac{\sigma_1^4}{f_1} + \cdots + a_k^2\frac{\sigma_k^4}{f_k}].\end{aligned}$$

Vi har her benyttet, at $\mathrm{V}(\chi^2(f)) = 2f$, d.v.s.

$$\mathrm{V}(\sigma^2\chi^2(f)/f) = 2\frac{\sigma^4}{f}.$$

Hvis nu MS var $\sigma^2\chi^2(f)/f$-fordelt, ville

$$f = \frac{2\,\mathrm{E}^2(MS)}{\mathrm{V}(MS)}.$$

Da nu $\hat{\sigma}_i^2 = \mathrm{MS}_i$, finder vi estimatet

$$f = \frac{[a_1MS_1 + \cdots + a_kMS_k]^2}{a_1^2\frac{MS_1^2}{f_1} + \ldots + a_k^2\frac{MS_k^2}{f_k}}.$$

Vi foretager nu testningen, som om MS var $\sigma^2\chi^2(\hat{f})/\hat{f}$-fordelt, og danner de relevante F-størrelser m.v.

Indholdet i ovenstående metode er, at vi opfatter MS som følgende den $\sigma^2\chi^2(f)/f$-fordeling, der har samme middelværdi og varians som MS. Hvis alle a_i'er er positive, vides dette at være en fremragende approximation til den eksakte fordeling. Hvis nogle a_i'er er negative, kender forfatteren ingen resultater angående kvaliteten af approximationen.

Inden vi giver et illustrativt eksempel, skal vi gøre opmærksom på, at der optræder en række variansanalysemodeller i litteraturen, der nok minder meget om de her omtalte, men som alligevel afviger på væsentlige punkter.

Vi præciserer, at metodernes gyldighed forudsætter, at der er tale om fuldstændige, balancerede faktorforsøg, hvor faktorer er krydsede eller hierarkiske, og alle involverede, stokastiske variable er forudsat **uafhængige**.

Hvis man ønsker at betragte afhængige **vekselvirknningsvariable** nemlig f.eks. T_{ij}'er, der tilfredsstiller

$$\sum_i T_{ij} = 0$$

i modellen

$$X_{ij\nu} = \mu + \alpha_i + B_j + T_{ij} + Z_{\nu(ij)},$$

må man ændre på reglerne for udfyldelse af "skemaet" p. 256. Fase 3 ændres til

3a. På de resterende pladser skrives 1, hvis **faktoren** er stokastisk, og 0, hvis den er deterministisk.

Ændringerne i de præcise formler for δ'erne er åbenbare. Vendingen "blot en af D'erne (eller B'erne eller C'erne) er stokastiske" erstattes med "og F_i er stokastisk". Tilsvarende erstattes "og alle D'er (og B'er og C'er) er deterministiske" med "og F_i er deterministisk". Den præcise forudsætning, vi nu må gøre, er, at summen af vekselvirkningsvariable over indices, der hører til deterministiske faktorer, er 0, d.v.s. de er ikke længere uafhængige.

I de tilfælde, hvor der såvel er deterministiske som stokastiske faktorer, er de F-tests, man kommer frem til, kun approximative ([31] p. 270). Hvis kun en hovedeffekt er stokastisk, kan man dog opnå eksakte tests ved hjælp af Hotellings T^2-størrelse, jvf. [31] p. 270 og 288.

Hvis der ønskes taget hensyn til, at en faktor muligvis kun kan forekomme i endeligt mange niveauer, e.g. H, og vi i vores forsøg tilfældigt har udvalgt h blandt disse, er-

stattes 1-taller i det grundlæggende skema med størrelsen

$$1 - \frac{h}{H}.$$

Dette gælder dog ikke 1-taller affødt ved, at et index er hierarkisk underordnet andre. Sådanne 1-taller bevares. Der kan henvises til [16] p. 144-5 eller [4] p. 414.

Det forekommer forfatteren, at begrebet afhængige vekselvirkningsvariable giver en lidt mærkelig ad hoc forklaring af den datagenerende proces, men disse modeller har vundet stort indpas i litteraturen. Således fremføres de af [4] p. 414, [13] p. 177, [16] p. 144 og [31] p. 284, og af ovenstående nævner kun [16] p. 145 tilfældet med uafhængige vekselvirkninger, skønt disse modeller ofte må foretrækkes.

Det må dog indskærpes, at det er væsentligt at gøre sig klart, hvilken model man opererer med. Fremgangsmåden ved testningen kan være forskellig ved de to modeller (og selvfølgelig bliver tolkningen af resultater altid forskellig).

Beviser for, at skemaerne til brug for beregninger af forventede værdier af SAK/f'er er korrekt, må føres induktivt. Et sådant bevis er skitseret hos [7]. Også [33] skitserer beviser.

Slutteligen vil det vel være rimeligt kort at komme med nogle korte bemærkninger vedrørende variansanalysernes robusthed overfor afvigelser fra forudsætningerne.

Den grundlæggende antagelse er, at residualstørrelserne $Z_{ij\ldots\nu}$ er normalt fordelte, uafhængige og har samme varians σ^2. Det viser sig imidlertid, at normalitetsantagelsen ikke er særligt kritisk. Styrkefunktionen for testene er relativt ufølsom over for selv relativt store afvigelser i fordelingernes form.

Sådanne afvigelser kan imidlertid få "katastrofale" konsekvenser for et eventuelt test for, at varianserne er ens. Således er Bartletts test meget følsomt over for afvigelser fra normalitet. Der findes dog andre ikke helt så følsomme tests (se bind 1 p. 5.82).

Til gengæld viser det sig imidlertid, at kravet om varianshomogenitet ikke er helt så afgørende, som man kunne forestille sig, hvis forsøget er balanceret, i.e. hvis der er lige mange observationer i hver celle.

For en mere dybtgående diskussion må henvises til litteraturen, e.g. [31] p. 331.

Vi giver nu et illustrativt eksempel.

EKSEMPEL 5.7. Nedenstående data ([4]) giver summerne af tre gentagne bestemmelser af 700%-elasticitetsmodulet for emner fremstillet af gummi, stammende fra 4 forskellige ladninger rågummi.

For hver ladning har man efter en omhyggelig blandingsprocedure udvalgt to stikprøver af gummi fra hver af 4 baller rågummi.

hver stikprøve er dernæst delt i to sub-stikprøver. Disse blev dernæst brugt til at fremstille standardblandinger indeholdende henholdsvis 0.5% og 3.0% stearinsyre. Prøver fra hver af disse blandinger blev dernæst hærdet i 60 eller 120 minutter, hvorefter 700%-elasticitetsmodulet er bestemt.

Rækkefølgen, hvori disse operationer udføres, blev omhyggeligt randomiseret, således at et tidstrend ikke skulle kunne influere på de effekter, man var interesseret i at undersøge.

Vi ser, at der i alt optræder 5 faktorer, nemlig

$$\begin{array}{lcll} A & = & \text{stearinsyreniveau,} & m = 1,2 \\ B_k & = & \text{stikprøve-nr.,} & k = 1,2 \\ C_i & = & \text{ladnings-nr.,} & i = 1,\ldots,4 \\ D_\ell & = & \text{hærdningsniveau,} & \ell = 1,2 \\ E_j & = & \text{balle-nr.,} & j = 1,\ldots,4. \end{array}$$

Der foreligger sådan set også en gentagelseseffekt, men da vi kun har opgivet summerne af de tre gentagelsesmålinger og ikke også deres kvadratsummer, må vi se bort fra dem i dette eksempel.

Ved fastlæggelsen af en model vil det nu være hensigtsmæssigt at gå frem som følger.

i) Vi skal først afgøre, hvilke af ovenstående faktorer, der er krydsede og hvilke, der er nestede. Vi starter systematisk fra oven

$$\begin{array}{llll} A \times B, & A \times C, & A \times D, & A \times E, \\ B \subset c, & B \times D, & B \subset E, & \\ C \times D, & C \supset E, & & \\ D \times E. & & & \end{array}$$

ii) Ved fastlæggelse af middelværdistrukturen inddrager vi nu faktorerne 1 efter 1, og det giver, idet vi noterer hvert enkelt trin (vi har ikke taget stilling til hvilke faktorer, der skal opfattes deterministiske og hvilke, der skal opfattes stokastiske. Vi anvender derfor de samme symboler, som vi har anvendt for faktorerne, i.e. her: store bogstaver),

$$\begin{array}{ll} & A \\ + & B(EC) + AB(EC) \\ + & C + AC \\ + & D + AD + BD(EC) + ABD(EC) + CD + ACD \\ + & E(C) + AE(C) + DE(C) + ADE(C). \end{array}$$

Vi bemærker, at når en ny faktor inddrages, skal man a) huske at angive alle faktorer, der er overordnet den pågældende faktor, i en parentes bag faktoren. Eksempelvis er den faktor, der inddrages som nr. 2, faktoren B. Ved at checke i listen ovenfor ser

		C_1				C_2				C_3				C_4			
		B_1		B_2		B_1		B_2		B_1		B_2		B_1		B_2	
		A_1	A_2	A_1	A_2	A_1	A_2	A_1	A_2	A_1	A_2	A_1	A_2	A_1	A_2	A_1	A_2
E_1	D_1	3810	6340	4080	6660	3004	5470	2833	6220	4210	7640	3910	6510	3810	8100	4570	8310
	D_2	6230	9050	6370	8650	5720	8390	5620	7300	6190	9720	6860	9960	8530	10210	7450	10650
E_2	D_1	4350	6450	3800	6920	3380	6330	3308	6280	3310	6040	3302	6670	4380	8010	4860	8100
	D_2	6250	9090	6700	9000	6060	8770	5820	9180	5660	8540	5690	8720	7130	10170	7350	9430
E_3	D_1	3690	6510	3590	6510	2758	5740	2945	6950	3445	6690	3273	5980	9160	7370	4670	7380
	D_2	6630	8680	6360	8440	5630	8200	5170	8310	6180	8060	5700	8210	7380	9560	7610	9400
E_4	D_1	4390	7470	4700	7210	3810	6830	3320	6020	2684	6010	3080	5630	5270	8010	4640	7540
	D_2	7170	9370	7460	8880	6280	9240	6170	8680	5790	7980	5180	7960	7350	10510	7590	10420

Tabel 5.2: Data til eksempel 5.7

vi, at B er underordnet såvel C som E. De skal derfor altid optræde i en parentes efter B. Derefter skal man b) anføre de led, der fremkommer ved at "parre" den sidst inddragne faktor med alle de led, der allerede forekommer. Herunder gøres igen brug af de under i) anførte relationer mellem faktorerne, idet man samtidigt erindrer, at hvis en faktorkombination allerede forekommer, skal den selvfølgelig ikke med igen, ligesom et "bogstav" udelades uden for en parentes, hvis det ved "parringen" optræder såvel inden i som uden for en parentes. E.g. har vi ved inddragelsen af $E(C)$ parrene

$$\begin{aligned} AC \quad &\& \quad E(C), \\ BD(EC) \quad &\& \quad E(C). \end{aligned}$$

Her giver det første anledning til

$$AE(C),$$

fordi A og E er krydsede. Det andet par giver ikke anledning til noget nyt led, da alle bogstaver i $E(C)$ allerede forekommer i $BD(EC)$.

iii) Vi skal nu afgøre, hvilke af de anførte effekter, der er tilfældige og hvilke, der er systematiske.

Faktorerne, der angiver stearinsyreniveauet og hærdningsniveauet (i.e. A og D) må anses for at være systematiske. Tilsvarende må balle-nr. og stikprøve-nr. (faktorerne E og B) anses for at være tilfældige. Det er lidt mere tvivlsomt, hvad ladningsnr. skal sættes til. Hvis der f.eks. er tale om 9 leverancer fra hvert af de 4 største producentlande, ville det nok være rimeligt at regne faktoren for systematisk. Hvis der imidlertid er tale om tilfældige ladninger indkøbt på en auktion, vil man regne den for tilfældig. Der foreligger ikke yderligere informationer, så vi må træffe et valg. Vi anser den sidste mulighed for at være den mest rimelige, således at vi regner C for tilfældig.

Nu regnes alle de effekter, hvori der **kun** indgår systematiske faktorer, for systematiske og resten for tilfældige.

Dette giver anledning til den endelige middelværdiopspaltning

$$\begin{aligned} X_{mki\ell j\nu} \quad = \quad & \mu + a_m + B(EC)_{k(ji)} + aB(EC)_{mk(ji)} \\ + \quad & C_i + aC_{mi} + d_\ell + ad_{m\ell} + Bd(EC)_{k\ell(ji)} \\ + \quad & aBd(EC)_{mk\ell(ji)} + Cd_{i\ell} + aCd_{mi\ell} + E(C)_{j(i)} \\ + \quad & aE(C)_{mj(i)} + dE(C)_{\ell j(i)} + adE(C)_{m\ell j(i)} \\ + \quad & Z_{\nu(mki\ell j)}, \end{aligned}$$

hvor gentagelsesindex ν her kun antager 1 værdi.

Her gælder så de sædvanlige forudsætninger om de enkelte led.

iv) For at bestemme de forventede værdier af de SAK'er, der svarer til ovennævnte effekter, udfylder vi nu det p. 256 anførte skema.

Tallene i parentes under de enkelte faktorer i tabellen p. 264 angiver antallet af niveauer for den pågældende faktor. Et d i en parentes bag en effekt angiver, at effekten er **deterministisk**, og et t, at den er **tilfældig**. Søjlen med R svarer til gentagelseseffekten. (Vi har ingen muligheder for at finde den, men det udelukker selvsagt ikke, at den kan være til stede.)

Ved hjælp af de p. 257 anførte regler finder vi dernæst følgende skema (p. 264) over de forventede værdier af middelkvadratafvigelsessummerne.

I skemaet er anvendt betegnelsen ϕ^2_{AD} i stedet for σ^2_{AD} for at angive, at effekten AD er systematisk. Analogt med effekterne A og D. Som fodtegn anvender vi de under ii) anførte betegnelser for effekterne.

Effekt		Faktor					
		C (4)	E (4)	B (2)	D (2)	A (2)	R (3)
C	(t)	1	4	2	2	2	3
D	(d)	4	4	2	0	2	3
A	(d)	4	4	2	2	0	3
$E(C)$	(t)	1	1	2	2	2	3
CD	(t)	1	4	2	1	2	3
AC	(t)	1	4	2	2	1	3
AD	(d)	4	4	2	0	0	3
$B(CE)$	(t)	1	1	1	2	2	3
$DE(C)$	(t)	1	1	2	1	2	3
$AE(C)$	(t)	1	1	2	2	1	3
ACD	(t)	1	4	2	1	1	3
$BD(CE)$	(t)	1	1	1	1	2	3
$AH(CE)$	(t)	1	1	1	2	1	3
$ADE(C)$	(t)	1	1	2	1	1	3
$ABD(CE)$	(t)	1	1	1	1	1	3
$R(ABDCE)$	(t)	1	1	1	1	1	1

Tabel til bestemmelse af forventede værdier af kvadratafvigelsessummer.

Ved hjælp af et standardprogram (ANOVA se p. 271) kan vi finde kvadratsummer svarende til en fuldstændigt krydset 5-faktor variansanalyse. Resultatet af en sådan kørsel er vist p. 265. Vi finder nu variansanalyseskemaet svarende til den her foreliggende struktur ved at addere de forskellige SAK-værdier efter de regler, der er anført p. 255.

Dette er gjort i skemaet p. 266. Nu må det imidlertid erindre at vi har foretaget beregningerne på summerne af de gentagne bestemmelser og ikke på gennemsnittene. Dette bevirker selvsagt, at de anførte SAK'er er ganget med en faktor 3 i forhold til det, de bør være. Derfor optræder der en $\mathrm{SAK} = \mathrm{SAK}'/3$-søjle. (Vi har udeladt $\sum_\nu$. Der er de $9/3 = 3$ for store).

Effekt	E(SAK/f)
C	$96\sigma^2_C + 24\sigma^2_{E(C)} + 48\sigma^2_{CD} + 48\sigma^2_{AC} + 12\sigma^2_{B(CE)} + 12\sigma^2_{DE(C)} + 12\sigma^2_{AE(C)} + 24\sigma^2_{ACD} + 6\sigma^2_{BD(CE)} + 6\sigma^2_{AB(CE)} + 6\sigma^2_{ADE(C)} + 3\sigma^2_{ABD(CE)} + \sigma^2$
D	$192\phi^2_D + 48\sigma^2_{CD} + 12\sigma^2_{DE(C)} + 24\sigma^2_{ACD} + 6\sigma^2_{BD(CE)} + 6\sigma^2_{ADE(C)} + 3\sigma^2_{ABD(CE)} + \sigma^2$
A	$192\phi^2_A + 48\sigma^2_{AC} + 12\sigma^2_{AE(C)} + 24\sigma^2_{ACD} + 6\sigma^2_{AB(CE)} + 6\sigma^2_{ADE(C)} + 3\sigma^2_{ABD(CE)} + \sigma^2$
$E(C)$	$24\sigma^2_{E(C)} + 12\sigma^2_{B(CE)} + 12\sigma^2_{DE(C)} + 12\sigma^2_{AE(C)} + 6\sigma^2_{BD(CE)} + 6\sigma^2_{AB(CE)} + 6\sigma^2_{ADE(C)} + 3\sigma^2_{ABD(CE)} + \sigma^2$
CD	$48\sigma^2_{CD} + 12\sigma^2_{DE(C)} + 24\sigma^2_{ACD} + 6\sigma^2_{BD(CE)} + 6\sigma^2_{ADE(C)} + 3\sigma_{ABD(CE)} + \sigma^2$
AC	$48\sigma^2_{AC} + 12\sigma^2_{AE(C)} + 24\sigma^2_{ACD} + 6\sigma^2_{AB(CE)} + 6\sigma^2_{ADE(C)} + 3\sigma^2_{ABD(CE)} + \sigma^2$
AD	$96\phi^2_{AD} + 24\sigma^2_{ACD} + 6\sigma^2_{ADE(C)} + 3\sigma^2_{ABD(CE)} + \sigma^2$
$B(CE)$	$12\sigma^2_{B(CE)} + 6\sigma^2_{BD(CE)} + 6\sigma^2_{AB(CE)} + 3\sigma^2_{ABD(CE)} + \sigma^2$
$DE(C)$	$12\sigma^2_{DE(C)} + 6\sigma^2_{BD(CE)} + 6\sigma^2_{ADE(C)} + 3\sigma^2_{ABD(CE)} + \sigma^2$
$AE(C)$	$12\sigma^2_{AE(C)} + 6\sigma^2_{AB(CE)} + 6\sigma^2_{ADE(C)} + 3\sigma^2_{ABD(CE)} + \sigma^2$
ACD	$24\sigma^2_{ACD} + 6\sigma^2_{ADE(C)} + 3\sigma^2_{ABD(CE)} + \sigma^2$
$BD(CE)$	$6\sigma^2_{BD(CE)} + 3\sigma^2_{ABD(CE)} + \sigma^2$
$AB(CE)$	$6\sigma^2_{AB(CE)} + 3\sigma^2_{ABD(CE)} + \sigma^2$
$ADE(C)$	$6\sigma^2_{ADE(C)} + 3\sigma^2_{ABD(CE)} + \sigma^2$
$ABD(CE)$	$3\sigma^2_{ABD(CE)} + \sigma^2$

Source of variation	Sums of squares	Degrees of freedom
A	247 331 000	1
B	80 501	1
AB	27 907	1
C	47 874 590	3
AC	1 010 358	3
BC	61 802	3
ABC	164 078	3
D	191 805 600	1
AD	1 802 150	1
BD	113 407	1
ABD	5 126	1
CD	282 293	3
ACD	349 272	3
BCD	356 120	3
$ABCD$	526 729	3
E	2 658 565	3
AE	356 797	3
BE	415 995	3
ABE	236 395	3
CE	11 951 750	9
ACE	1 207 328	9
BCE	708 651	9
$ABCE$	1 066 008	9
DE	251 008	3
ADE	357 499	3
BDE	174 139	3
$ABDE$	142 968	3
CDE	884 268	9
$ACDE$	1 408 724	9
$BCDE$	1 827 469	9
$ABCDE$	1 278 840	9
Total	516 714 200	127

Kvadratafvigelsessummer fundet med ANOVA. Grundet afrundingsfejl summerer de anførte SAK'er ikke op til den givne total.

Variationskilde	Faktorkombination	SAK′	SAK = SAK′/3	f	SAK/f
C	C	47 874 590	15 958 197	3	5 319 399
D	D	191 805 600	63 935 200	1	63 935 200
A	A	247 331 000	82 443 667	1	82 443 667
$E(C)$	$E + CE$	14 610 315	4 870 105	12	405 842
CD	CD	282 293	94 098	3	31 366
AC	AC	1 010 358	336 786	3	112 262
AD	AD	1 802 150	600 717	1	600 717
$B(CE)$	$B + BC + BE + BCE$	1 266 949	422 316	16	26 395
$DE(C)$	$DE + CDE$	1 135 277	378 426	12	31 536
$AE(C)$	$AE + ACE$	1 564 125	521 375	12	43 448
ACD	ACD	349 272	116 424	3	38 808
$BD(CE)$	$BD + BCD + BDE + BCDE$	2 471 134	823 711	16	51 482
$AB(CE)$	$AB + ABC + ABE + ABCE$	1 494 387	498 129	16	31 133
$ADE(C)$	$ADE + ACDE$	1 766 223	588 741	12	49 062
$ABD(EC)$	$ABD + ABDE + ABCD + ABCDE$	1 953 663	651 221	16	40 701
Total		516 714 200	172 238 067	127	1 356 205

Ved hjælp af variansanalyseskemaet og skemaet over de forventede værdier af SAK/f kan vi nu teste en række hypoteser om effekterne og varianskomponenterne.

Eksempelvis har vi, at

$$\frac{49062}{40701} = 1.21 < \mathrm{F}(12,16)_{0.90} = 1.99,$$

hvorfor vi i hvert fald for alle $\alpha > 10\%$ vil antage hypotesen, at

$$\sigma^2_{ADE(C)} = 0.$$

Tilsvarende ses, at vi kan antage

$$\sigma^2_{AB(CE)} = 0$$
$$\sigma^2_{BD(CE)} = 0.$$

Når vi nu vil undersøge, om det kan antages, at $\sigma^2_{ACD} = 0$, kan vi enten betragte teststørrelsen

$$\frac{\mathrm{SAK}_{ACD}/3}{\mathrm{SAK}_{ADE(C)}/12},$$

som er $\mathrm{F}(3,12)$-fordelt under H_0, eller vi kan benytte os af, at vi har fået accepteret, at $\sigma^2_{ADE(C)} = 0$, og så betragte

$$\frac{\mathrm{SAK}_{ACD}/3}{\text{skøn over } \{\sigma^2 + 3\sigma^2_{ABD(CE)}\}}.$$

Som nævner kan vi da, stadig under forudsætning af at

$$\sigma^2_{AB(CE)} = 0,$$
$$\sigma^2_{BD(CE)} = 0,$$

og

$$\sigma^2_{ADE(C)} = 0,$$

bruge

$$S_1^2 = \frac{\mathrm{SAK}_{ABD(CE)} + \mathrm{SAK}_{ADE(C)} + \mathrm{SAK}_{AB(CE)} + \mathrm{SAK}_{BD(CE)}}{16 + 12 + 16 + 16}.$$

Teststørrelsen bliver da $\mathrm{F}(3,60)$-fordelt under H_0. Fordelen ved denne fremgangsmåde er åbenbart, at vi får flere frihedsgrader i nævneren, d.v.s. vores skøn over den varians,

vi skal sammenligne med, bliver mere præcis. Ulempen er selvsagt, at selv om vi har fået accepteret, at $\sigma^2_{AB(CE)} = 0$ etc., er dette jo ikke nødvendigvis rigtigt, således at teststørrelsen ikke har den fordeling, som vi regner med.

For at løse problemet med, om man skal "poole" disse varianser, kan man anvende en håndregel som f.eks. kun at poole varianser, hvis F-teststørrelsen er mindre end 2. Dette er selvsagt en vilkårlig regel, men den nyder en vis udbredelse i praksis (cf. [16] vol. II).

Anvendes denne fremgangsmåde, fås teststørrelsen

$$\frac{38808}{2561802/60} = 0.91 \simeq \mathrm{F}(3,60)_{50\%},$$

og det ses, at også hypotesen

$$\sigma^2_{ACD} = 0$$

accepteres.

Fortsættes på denne måde, ses, at vi må antage

$$\begin{array}{lrcl} \exists m,\ell & : & ad_{m\ell} & \neq & 0 \\ & & \sigma^2_{AC} & \neq & 0 \\ & & \sigma^2_{E(C)} & \neq & 0. \end{array}$$

Ved en undersøgelse af, om alle a_m'er er 0, ses, at vi kan anvende teststørrelsen

$$\frac{\mathrm{SAK}_A/1}{\mathrm{SAK}_{AC}/3} = \frac{82443667}{112262} = 734.4,$$

som skal sammenlignes med en $\mathrm{F}(1,3)$-fraktil. Det fremgår, at vi på alle rimelige niveauer må antage, at der eksisterer a_m'er forskellige fra 0.

Når vi skal undersøge d_ℓ'erne, kan vi anvende teststørrelsen

$$\frac{\mathrm{SAK}_D/1}{S_1^2} = \frac{63935200}{2561802/60} = 1497.4,$$

og det ses igen, at vi på alle rimelige niveauer får forkastet hypotesen, at alle d_ℓ'er er lig 0.

Endelig mangler vi at undersøge σ^2_C. Hvis vi får ud fra alle de hypoteser, vi indtil nu har fået bekræftet, ses, at

$$\mathrm{E}(\mathrm{SAK}_C/f_C) = 96\sigma^2_C + 24\sigma^2_{E(C)} + 48\sigma^2_{AC} + 3\sigma^2_{ABD(CE)} + \sigma^2,$$

og

$$\begin{aligned} \mathrm{E}(\mathrm{SAK}_{E(C)}/f_{E(C)} &= 24\sigma_{E(C)} + 3\sigma_{ABD(CE)} + \sigma^2 \\ \mathrm{E}(\mathrm{SAK}_{AC}/f_{AC}) &= 48\sigma_{AC} + 3\sigma_{ABD(CE)} + \sigma^2 \end{aligned}$$

Det er derfor ikke umiddelbart muligt at danne en teststørrelse, så vi må bruge et approximativt test, som angivet p. 257.

Vi finder, at

$$\begin{aligned} &\mathrm{E}\{\mathrm{SAK}_{E(C)}/f_{E(C)} + \mathrm{SAK}_{AC}/f_{AC} - S_1^2\} \\ &= \mathrm{E}\{S_{E(C)}^2 + S_{AC}^2 - S_1^2\} \\ &= 24\sigma_{E(C)}^2 + 48\sigma_{AC}^2 + 3\sigma_{ABD(CE)}^2 + \sigma^2. \end{aligned}$$

Vi kan derfor bruge størrelsen i den krøllede parentes som nævner ved et approximativt F-test. Idet

$$S_1^2 = 2\,561\,802/60 = 42\,697,$$

estimerer vi antallet af frihedsgrader til (jvf. p. 258)

$$\begin{aligned} \hat{f} &= \frac{[405\,842 + 112\,262 - 42\,697]^2}{\frac{405\,842^2}{12} + \frac{112\,262^2}{3} + \frac{42\,697^2}{60}} \\ &= 12.59, \end{aligned}$$

og selve variansskønnet er

$$S^2 = 405\,842 + 112\,262 - 42\,697 = 475\,407.$$

Teststørrelsen bliver derfor

$$\frac{5\,319\,399}{475\,407} = 11.19,$$

som skal sammenlignes med en F(3, 12.6)-fordeling. Det ses, at resultatet af dette approximative F-test bliver, at vi må antage

$$\sigma_C^2 \neq 0.$$

Som resultat af denne analyse må vi derfor antage, at der er en effekt af ladningsnr., af hærdningsniveau, af stearinsyreniveau og af ballenr.inden for ladningsniveau. Endvidere er der vekselvirkning mellem stearinsyreniveau og hærdningsniveau og mellem stearinsyreniveau og ladningsnr.

En viderefortolkning af disse resultater og en angivelse af estimater overlades til læseren. ♦

Kapitel 6

Test 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 $\mathbf{X}_1, \ldots, \mathbf{X}_n$, hvor

$$\mathbf{X}_i \in \mathrm{N}_p(\mu, \Sigma),$$

d.v.s p-dimensionalt normalt fordelt med middelværdivektor μ og dispersionsmatrix Σ. Det forudsættes at Σ er regulær og ukendt. Vi ønsker at teste en hypotese om, at middelværdivektoren μ er lig en given vektor μ_0 mod alle alternativer, i.e.

$$H_0 : \mu = \mu_0 \quad \text{mod} \quad H_1 : \mu \neq \mu_0$$

Vi repeterer først nogle resultater om estimationerne. Fra sætning 2.27 p. 103 har vi følgende resultater om den empiriske middelværdivektor $\bar{\mathbf{X}}$ og den empiriske disper-

sionsmatrix $\mathbf{S}$

$$\bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i \qquad \in \mathrm{N}_p(\mu, \tfrac{1}{n}\Sigma)$$
$$\mathbf{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})' \quad \in \mathrm{W}(n-1, \tfrac{1}{n-1}\Sigma)$$

$\bar{\mathbf{X}}$ og $\mathbf{S}$ er stokastisk uafhængige.

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

LEMMA 6.1. Lad $\mathbf{Y}$ være en p-dimensional stokastisk variabel og lad U være en $p \times p$ stokastisk matrix med

$$\begin{aligned} \mathbf{Y} &\in \mathrm{N}_p(\mu, \Sigma) \\ m\mathbf{U} &\in \mathrm{W}(m, \Sigma), \end{aligned}$$

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

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

Da gælder

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

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

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

SÆTNING 6.1. ▲

BEVIS 6.1. Forbigås. Se e.g. [3], p. 106 ■

Vi har nu følgende hovedresultat

SÆTNING 6.2. Vi anvender betegnelsen

$$T^2 = n(\bar{\mathbf{X}} - \mu_0)'\mathbf{S}^{-1}(\bar{\mathbf{X}} - \mu_0),$$

hvor $\bar{\mathbf{X}}$, μ_0 og $\mathbf{S}$ 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

$$C = \{\mathbf{x}_1, \ldots, \mathbf{x}_n | \frac{n-p}{(n-1)p} t^2 > \mathrm{F}(p, n-p)_{1-\alpha}\},$$

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

BEVIS 6.2. Af lemma 6.1 følger, at

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

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 kvortienttestet følger ved direkte regning bl.a. under benyttelse af sætning 1.2. ■

BEMÆRKNING 6.1. Størrelsen T^2 kaldes ofte Hotelling's T^2 efter Harold Hotelling, der først betragtede denne teststørrelse. ▼

BEMÆRKNING 6.2. 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 studentfordelt variabel $\mathrm{t}(\nu)$ er $\mathrm{F}(1, \nu)$-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

$$T^2 = \frac{\det[\mathbf{S} + n(\bar{\mathbf{X}} - \mu_0)(\bar{\mathbf{X}} - \mu_0)']}{\det[\mathbf{S}]} - 1$$

▲

BEVIS 6.3. Forbigås. Rent teknisk og følger ved anvendelse af sætning 1.2 p. 17 på matricen

$$\begin{bmatrix} -1 & \sqrt{n}(\bar{\mathbf{X}} - \mu_0)' \\ \sqrt{n}(\bar{\mathbf{X}} - \mu_0) & \mathbf{S} \end{bmatrix}$$

■

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

$$\mu_0 = \begin{pmatrix} 22.10 \\ 7.40 \end{pmatrix}.$$

Det synes rimeligt, at anvende Hotelling's T^2 til at afgøre ovenstående spørgsmål. Kaldes observationerne $0\mathbf{x}_1, \ldots, \mathbf{x}_7$, finder vi

$$\begin{aligned} \bar{\mathbf{x}} &= \begin{pmatrix} 19.66 \\ 5.31 \end{pmatrix}, \\ \mathbf{s} &= \begin{pmatrix} 1.1795 & -0.3076 \\ -0.3076 & 0.8681 \end{pmatrix}. \end{aligned}$$

Da

$$\bar{\mathbf{x}} - \mu_0 = \begin{pmatrix} -2.44 \\ -2.09 \end{pmatrix},$$

er

$$n(\bar{\mathbf{x}} - \mu_0)(\bar{\mathbf{x}} - \mu_0)' = \begin{pmatrix} 41.68 & 35.70 \\ 35.70 & 30.58 \end{pmatrix},$$

og

$$\mathbf{s} + n(\bar{\mathbf{x}} - \mu_0)(\bar{\mathbf{x}} - \mu_0)' = \begin{pmatrix} 42.86 & 35.39 \\ 35.39 & 31.45 \end{pmatrix}.$$

Følgelig er

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

F-teststørrelsen er

$$\frac{7-2}{6 \cdot 2} t^2 = 42.8 > \mathrm{F}(2.5)_{0.999} = 37.1,$$

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 konfidensområdet for μ. Vi har med den sædvanlige notation

SÆTNING 6.4. Et $(1-\alpha)$-konfidensområde for forventningen $\mathrm{E}(\mathbf{X})$ er

$$\{\mu | n(\bar{\mathbf{x}} - \mu)'\mathbf{s}^{-1}(\bar{\mathbf{x}} - \mu) \leq \frac{(n-1)p}{n-p}\mathrm{F}(p, n-p)_{1-\alpha}\},$$

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

BEVIS 6.4. Triviel følge af definitionen på et konfidensområde og sætning 6.2 ■

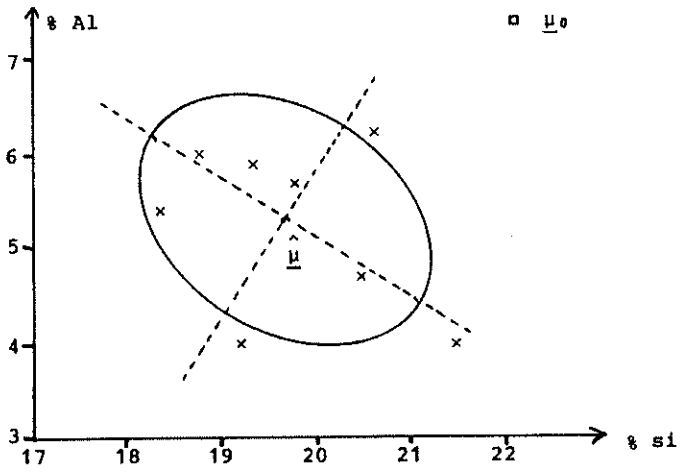
Vi fortsætter nu eksempel 6.1 i nedenstående

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

$$7(19.66 - \mu_1, 5.31 - \mu_2)\mathbf{s}^{-1} \begin{pmatrix} 19.66 - \mu_1 \\ 5.31 - \mu_2 \end{pmatrix} = \frac{12}{5}\mathrm{F}(2.5)_{0.95}$$

eller

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



Figur 6.1: Observationer og konfidensområde.

Vi finder

$$\mathbf{s}^{-1} = \begin{pmatrix} 0.9341 & 0.3310 \\ 0.3310 & 1.2692 \end{pmatrix}$$

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 $\bar{\mathbf{x}}$ og med ovenstående vektorer som enhedsvektorer har ellipsen ligningen

$$1.4727y_1^2 + 0.7307y_2^2 = 1.9851$$

eller

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

I figur 6.1 er konfidensområdet og observationerne anført. Desuden er $\mu_0 = (22.10,\ 7.40)'$ anført. Det ses, at dette punkt ligger uden for konfidensområdet i overensstemmelse med, at hypotesen $\mu = \mu_0$ mod $\mu \neq \mu_0$ 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 $\mathbf{X}_1, \ldots, \mathbf{X}_n$ og $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$, hvor

$$\begin{aligned} \mathbf{X}_i &\in \mathrm{N}_p(\mu, \boldsymbol{\Sigma}) \\ \mathbf{Y}_i &\in \mathrm{N}_p(\nu, \boldsymbol{\Sigma}), \end{aligned}$$

og vi ønsker at teste

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

Vi anvender betegnelserne

$$\begin{aligned} \bar{\mathbf{X}} &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i \\ \bar{\mathbf{Y}} &= \frac{1}{m}\sum_{i=1}^{m}\mathbf{Y}_i \\ \mathbf{S}_1 &= \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})' \\ \mathbf{S}_2 &= \frac{1}{m-1}\sum_{i=1}^{m}(\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})' \\ \mathbf{S} &= \frac{(n-1)\mathbf{S}_1 + (m-1)\mathbf{S}_2}{n+m-2} \end{aligned}$$

Ifølge sætning 2.27 og sætning 2.26 har vi

$$\begin{aligned} \bar{\mathbf{X}} &\in \mathrm{N}_p(\mu, \frac{1}{n}\boldsymbol{\Sigma}) \\ \bar{\mathbf{Y}} &\in \mathrm{N}_p(\nu, \frac{1}{m}\boldsymbol{\Sigma}) \\ \mathbf{S} &\in \mathrm{W}(n+m-2, \frac{1}{n+m-2}\boldsymbol{\Sigma}). \end{aligned}$$

Vi formulerer nu hovedresultatet om testning af H_0 mod H_1 i

SÆTNING 6.5. Vi anvender de samme betegnelser som givet ovenfor. Vi sætter

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

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

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

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

BEVIS 6.5. Af lemma 6.1 og ovenfor anførte relationer følger, at

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

og heraf følger resultatet umiddelbart. ■

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 $\mu-\nu=\delta_o$. Da er et $(1-\alpha)$-konfidensområde for δ_o lig

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

▲

BEVIS 6.6. Direkte følge af definitionen på et konfidensområde og sætning 6.5. ■

BEMÆRKNING 6.3. Konfidensområdet er en ellipsoide med centrum i $\bar{\mathbf{x}}-\bar{\mathbf{y}}$ og hovedakser bestemt af egenvektorerne i $\mathbf{s}^{-1}$. ▼

BEMÆRKNING 6.4. De anførte testresultater og konfidensintervaller kræver som anført, at dispersionsmatricerne for $\mathbf{X}$- og for $\mathbf{Y}$-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 [3], 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^{2}hud i 3 timer,

iii) middeltemperatur i °C. Denne temperatur fås ved at måle hudtemperatur 14 forskellige steder hvert minut i 5 minutter (samme steder hver gang). Middeltemperaturen er således et gennemsnit af i alt $14 \times 5 = 70$ målinger,

på 16 mænd og 16 kvinder. Resultatet af forsøget er givet i tabellen p. 282.

Vi opfatter disse tal som realisationer af stokastiske variable

$$\mathbf{X}_1, \ldots, \mathbf{X}_{16} \quad \text{og} \quad \mathbf{Y}_1, \ldots, \mathbf{Y}_{16}.$$

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

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

og

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

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

Estimaterne for μ og ν er de empiriske middelværdier, d.v.s.

$$\hat{\mu} = \bar{\mathbf{x}} = \begin{pmatrix} 179.7 \\ 24.5 \\ 33.6 \end{pmatrix}$$

og

$$\hat{\nu} = \bar{\mathbf{y}} = \begin{pmatrix} 166.1 \\ 20.5 \\ 33.4 \end{pmatrix}.$$

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

Person nr.	Højde i cm	Fordampningstab i g/m^2 hud	Middeltemperatur i °C
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	39.2
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	33.5
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	39.0
16	180	12.6	33.5

Tabel 6.1: Data fra indeklimaforsøg på laboratoriet for Varme- og Klimateknik, DTH.

Vi finder med de i sætning 6.5 valgte betegnelser

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

og dermed

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

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 $\mu = \nu$ 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 teststørrelsen

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

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

6.2 Den flerdimensionale generelle lineære model

I de foregående afsnit har vi betragtet en- og tostikprøvesituationen 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 regressions- og variansanalyser af flerdimensionale variable.

Vi betragter indbyrdes uafhængige observationer $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$,

$$\mathbf{Y}_i \in \mathrm{N}_p(\mu_i, \boldsymbol{\Sigma}).$$

Dispersionsmatricen $\boldsymbol{\Sigma}$ (og middelværdivektorerne μ_i) antages ukendte. Vi ordner observationerne i en $n \times p$ datamatrix

$$\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1' \\ \vdots \\ \mathbf{Y}_n' \end{bmatrix} = \begin{bmatrix} Y_{11} & \cdots & Y_{1p} \\ \vdots & & \vdots \\ Y_{n1} & \cdots & Y_{np} \end{bmatrix}.$$

Her repræsenterer de enkelte rækker altså f.eks. 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 μ_i kan skrives som kendte lineære funktioner af andre (og færre) ukendte parametre θ, d.v.s.

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

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

$$\mathbf{Y}_{j|} = \begin{bmatrix} Y_{1j} \\ \vdots \\ Y_{nj} \end{bmatrix},$$

kan vi skrive

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

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

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

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

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

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

hvilket er en analog til den endimensionale regressionsmodel.

Ordnes observationerne i en søjlevektor

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

får vi af sætning 2.7, p. 63, at

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

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

Et første problem er at estimere θ. Der gælder

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

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

▲

BEVIS 6.7. Forbigås. Se f.eks. [3]. ■

BEMÆRKNING 6.5. Vi ser, at

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

d.v.s. estimatet for den j'te søjle i θ er simpelt hen lig det resultat, vi får ved kun at betragte den endimensionale generelle lineære model for den j'te "egenskab". ▼

BEMÆRKNING 6.6. Hvis observationerne ikke er normalt fordelte, vil man stadig kunne bruge det anførte skøn $\hat{\theta}$, 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

$$M = (\mathbf{Y} - \mathbf{x}\theta)'(\mathbf{Y} - \mathbf{x}\theta) - (\mathbf{Y} - \mathbf{x}\hat{\theta})'(\mathbf{Y} - \mathbf{x}\hat{\theta})$$

er positiv semidefinit. Dette medfører, at

$$\mathrm{ch}_i(\mathbf{Y}-\mathbf{x}\theta)'(\mathbf{Y}-\mathbf{x}\theta) \geq \mathrm{ch}_i(\mathbf{Y}-\mathbf{x}\hat{\theta})'(\mathbf{Y}-\mathbf{x}\hat{\theta}),$$

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

$$\det(\mathbf{Y}-\mathbf{x}\theta)'(\mathbf{Y}-\mathbf{x}\theta)$$

og

$$\mathrm{tr}(\mathbf{Y}-\mathbf{x}\theta)'(\mathbf{Y}-\mathbf{x}\theta).$$

▼

BEMÆRKNING 6.7. Vi har ovenfor stiltiende forudsat, at $\mathbf{x}'\mathbf{x}$ har fuld rang, d.v.s. at $\mathrm{rg}(\mathbf{x}) = k < n$. 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 $\hat{\theta}$ vender vi os mod estimation af $\boldsymbol{\Sigma}$.

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

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}^* &= \frac{1}{n}\sum_{i=1}^{n}(\mathbf{Y}_i-\hat{\theta}'\mathbf{x}_i)(\mathbf{Y}_i-\hat{\theta}'\mathbf{x}_i)' \\ &= \frac{1}{n}(\mathbf{Y}-\mathbf{x}\hat{\theta})'(\mathbf{Y}-\mathbf{x}\hat{\theta}) \\ &= \frac{1}{n}[\mathbf{Y}'\mathbf{Y}-(\mathbf{x}\hat{\theta})'(\mathbf{x}\hat{\theta})].\end{aligned}$$

Det (i,j)'te element kan også skrives

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

▲

BEVIS 6.8. De mange identiteter mellem $\hat{\boldsymbol{\Sigma}}$'s elementer fremgå¨r ved simple matrix-manipulationer. For selve resultatet henvises til [3]. ■

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

$$\tilde{\theta} = \mathrm{vc}(\theta) = \begin{bmatrix} \theta_{|1} \\ \vdots \\ \theta_{|p} \end{bmatrix}$$

$$\hat{\tilde{\theta}} = \mathrm{vc}(\hat{\theta}) = \begin{bmatrix} \hat{\theta}_{|1} \\ \vdots \\ \hat{\theta}_{|p} \end{bmatrix}.$$

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

$$\hat{\tilde{\theta}} = \mathrm{vc}(\hat{\theta}) \in \mathrm{N}_{pk}(\tilde{\theta}, \boldsymbol{\Sigma} \otimes (\mathbf{x}'\mathbf{x})^{-1}),$$

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

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

Endelig er $\boldsymbol{\Sigma}^*$ og $\hat{\tilde{\theta}}$ og dermed $\boldsymbol{\Sigma}^*$ og $\hat{\theta}$ stokastisk uafhængige. ▲

BEVIS 6.9. Det er trivielt, at

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

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

Ydermere gælder

$$\mathrm{D}(\hat{\theta}_{|i}) = \sigma_{ii}(\mathbf{x}'\mathbf{x})^{-1}$$

og

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

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

Resultatet vedrørende fordelingen af $\hat{\boldsymbol{\Sigma}}^*$ og vedrørende uafhængigheden af $\hat{\theta}$ og $\hat{\boldsymbol{\Sigma}}^*$ er helt analoge til de tilsvarende endimensionale resultater, men vi vil ikke komme nærmere ind på dette her. Læseren henvises til f.eks. [3]. ■

Af sætningen fås umiddelbnart

KOROLLAR 6.1. Det centrale skøn for $\boldsymbol{\Sigma}$ er lig

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

BEVIS 6.10. Følger trivielt, når det erindres, at

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

■

Vi vender os dernæst mod testning af parametre i modellen.

Der gælder

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

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

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

$$\begin{aligned} \boldsymbol{\Delta} &= \mathbf{A}\hat{\theta}\,\mathbf{B}' - \mathbf{C} \\ \mathbf{R} &= n\hat{\boldsymbol{\Sigma}}^* = (\mathbf{Y} - \mathbf{x}\hat{\theta})'(\mathbf{Y} - \mathbf{x}\hat{\theta}) \end{aligned}$$

og

$$\begin{aligned} \mathbf{S} &= \mathbf{B}\mathbf{R}\mathbf{B}' \\ \mathbf{H} &= \boldsymbol{\Delta}'[\mathbf{A}(\mathbf{x}'\mathbf{x})^{-1}\mathbf{A}']^{-1}\boldsymbol{\Delta}. \end{aligned}$$

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

$$\{\mathbf{y} | \frac{\det(\mathbf{s})}{\det(\mathbf{s}+\mathbf{h})} \leq U(s, r, n-k)_\alpha\},$$

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

BEVIS 6.11. Forbigås. Det bygger essentielt på, at det kan vises, at $\mathbf{S}$ og $\mathbf{H}$ 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

$$u = \frac{\det(\mathbf{S})}{\det(\mathbf{S}+\mathbf{H})}$$

alene afhængig af s, r og $n-k$. 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 den F-fordeling i nedenstående ■

SÆTNING 6.11. Lad U være $U(p,q,r)$-fordelt og sæt

$$\begin{aligned} t &= \begin{cases} 1 & p^2+q^2=5 \\ \sqrt{\frac{p^2q^2-4}{p^2+q^2-5}} & p^2+q^2\neq 5 \end{cases} \\ v &= \frac{1}{2}(2r+q-p-1). \end{aligned}$$

Da er

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

approximativt fordelt som

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

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

BEVIS 6.12. Forbigås. ■

Vi skal nu illustrere de indførte begreber 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, 9 = stærkest
5: Blomstringstid	"	1 = seneste blomstring, 9 = 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 Plante-højde	7 Frø-vægt	8 Plante-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	3.99	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	0.987	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 (Væksttype) kun er svagt korreleret med de øvrige variable, hvorimod f.eks. variabel 2 og 3 (genvækst og krybeevne) samt (naturligvis) 7 og 9 (frøvægt og % frø) er stærkt korrelerede.

Vi er som nævnt specielt interesserde 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.

$$\begin{aligned} \mathrm{E}(Y_1) &= \sum_{i=1}^{k} \theta_{i1} x_i \\ \mathrm{E}(Y_2) &= \sum_{i=1}^{k} \theta_{i2} x_i \end{aligned}$$

hvor vi nu har anvendt variabelbetegnelserne

$$\begin{aligned} Y_1 &= \text{krybeevne} \\ Y_2 &= \text{vigør} \\ x_1 &= \text{væksttype} \\ x_2 &= \text{genvækst efter vinter} \\ x_3 &= \text{blomstringstid} \\ x_4 &= \text{plantehøjde} \\ x_5 &= \text{frøvægt} \\ x_6 &= \text{plantevægt} \\ x_7 &= \text{procent frø.} \end{aligned}$$

Der er her åbenbart tale om den flerdimensional generel lineær model. Sætter vi $\theta =$

(θ_{ij}), fås

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

Antages

$$\begin{pmatrix} Y_{1i} \\ Y_{2i} \end{pmatrix} \in \mathrm{N}(\mu_i, \Sigma),$$

bliver det centrale skøn over Σ lig

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

Matricen $(\mathbf{x}'\mathbf{x})^{-1}$ 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

$$\mathrm{D}(\hat{\hat{\theta}}) = \Sigma \otimes (\mathbf{x}'\mathbf{x})^{-1} = \begin{pmatrix} \sigma_{11}(\mathbf{x}'\mathbf{x})^{-1} & \sigma_{12}(\mathbf{x}'\mathbf{x})^{-1} \\ \sigma_{21}(\mathbf{x}'\mathbf{x})^{-1} & \sigma_{22}(\mathbf{x}'\mathbf{x})^{-1} \end{pmatrix},$$

og dermed f.eks.

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

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

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

og

$$\mathbf{C} = \begin{pmatrix} 0 & 0 \end{pmatrix}.$$

Da bliver nemlig

$$\mathbf{A}\theta\mathbf{B}' = \begin{pmatrix} \theta_{41} & \theta_{42} \end{pmatrix}.$$

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

$$F = 53.66$$

med frihedsgrader

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

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

$$\theta_1 = \begin{bmatrix} \theta_{51} & \theta_{52} \\ \theta_{61} & \theta_{62} \\ \theta_{71} & \theta_{72} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

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

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

og

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

thi da er

$$\mathbf{A}\,\theta\,\mathbf{B}' = \theta_1.$$

Med det før omtalte standardprogram findes

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

Der er således fremdeles klar signifikans.

Som det sidste eksempel betragtes hypotesen

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

mod alle alternativer. Denne bringes på standardformen ved at vælge

$$\begin{aligned} \mathbf{A} &= \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \\ \mathbf{B} &= \begin{pmatrix} 0 & 1 \end{pmatrix} \end{aligned}$$

og

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

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

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

Disse antages at være stokastisk uafhængige med

$$\mathbf{Y}_{ij} \in \mathrm{N}_p(\mu_i, \boldsymbol{\Sigma}), \qquad i = 1, \ldots, k\,; \quad j = 1, \ldots, n_i,$$

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

Vi ønsker at teste hypotesen

$$H_0 : \mu_1 = \cdots = \mu_k$$

mod

$$H_1 : \exists i, j(\mu_i \neq \mu_j).$$

Vi definerer helt analogt til ensidede variansanalyser "kvadratafvigelsesmatricerne"

$$\begin{aligned}
\mathbf{T} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mathbf{Y}_{ij} - \bar{\mathbf{Y}})(\mathbf{Y}_{ij} - \bar{\mathbf{Y}})' \\
\mathbf{W} &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mathbf{Y}_{ij} - \bar{\mathbf{Y}}_i)(\mathbf{Y}_{ij} - \bar{\mathbf{Y}}_i)' \\
\mathbf{B} &= \sum_{i=1}^{k} n_i(\bar{\mathbf{Y}}_i - \bar{\mathbf{Y}})(\bar{\mathbf{Y}}_i - \bar{\mathbf{Y}})'
\end{aligned}$$

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

$$\begin{aligned}
\bar{\mathbf{Y}}_i &= \frac{1}{n_i}\sum_{j=1}^{n_i}\mathbf{Y}_{ij} \\
\bar{\mathbf{Y}} &= \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\mathbf{Y}_{ij}.
\end{aligned}$$

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

$$\mathbf{T} = \mathbf{W} + \mathbf{B},$$

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 $\boldsymbol{\Sigma}$ kan anvende

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-k}\mathbf{W}.$$

Hvis hypotesen er sand, vil også $\mathbf{T}$ være proportional med et sådant skøn. Hvis hypotesen ikke er sand, vil $\mathbf{T}$ 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

$$\{\mathbf{y}_{11}, \ldots, \mathbf{y}_{kn_k} \left| \frac{\det(\mathbf{W})}{\det(\mathbf{t})} \leq U(p, k-1, n-k)_\alpha \right.\}.$$

▲

BEVIS 6.13. Forbigås. Følger ved valg af særlige $\mathbf{A}$, $\mathbf{B}$ og $\mathbf{C}$ matricer i sætning 6.10. ■

Som for endimensionale variansanalyser samles resultaterne i et Variansanalyseskema

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

Det er selvfølgelig muligt, ligesom det er gjort i kapitel 5, at bestemme forventede værdier af $\mathbf{B}$ og $\mathbf{T}$ 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

$$\begin{matrix} \mathbf{Y}_{11}, & \ldots, & \mathbf{Y}_{1m} \\ \vdots & & \vdots \\ \mathbf{Y}_{k1}, & \ldots, & \mathbf{Y}_{km} \end{matrix},$$

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

$$\mathrm{E}(\mathbf{Y}_{ij}) = \mu_{ij} = \mu + \alpha_i + \beta_j,$$

hvor parametrene α_i og β_j tilfredsstiller

$$\sum_i \alpha_i = \sum_j \beta_j = \mathbf{0}.$$

Vi ønsker at teste hypoteserne

$$H_0 : \alpha_1 = \cdots = \alpha_k = \mathbf{0}$$

mod

$$H_1 : \exists i(\alpha_i \neq \mathbf{0})$$

og

$$K_0 : \beta_1 = \cdots = \beta_m = \mathbf{0}$$

mod

$$K_1 : \exists j(\beta_j \neq \mathbf{0}).$$

Vi definerer helt analogt til den endimensionale variansanalyse matricerne

$$\begin{aligned}
\mathbf{T} &= \sum_{i=1}^{k}\sum_{j=1}^{m}(\mathbf{Y}_{ij} - \bar{\mathbf{Y}}_{..})(\mathbf{Y}_{ij} - \bar{\mathbf{Y}}_{..})' \\
\mathbf{Q}_1 &= \sum_{i=1}^{k}\sum_{j=1}^{m}(\mathbf{Y}_{ij} - \bar{\mathbf{Y}}_{i.} - \bar{\mathbf{Y}}_{.j} + \bar{\mathbf{Y}}_{..})(\mathbf{Y}_{ij} - \bar{\mathbf{Y}}_{i.} - \bar{\mathbf{Y}}_{.j} + \bar{\mathbf{Y}}_{..})' \\
\mathbf{Q}_2 &= m\sum_{i=1}^{k}(\bar{\mathbf{Y}}_{i.} - \bar{\mathbf{Y}}_{..})(\bar{\mathbf{Y}}_{i.} - \bar{\mathbf{Y}}_{..})' \\
\mathbf{Q}_3 &= k\sum_{j=1}^{m}(\bar{\mathbf{Y}}_{.j} - \bar{\mathbf{Y}}_{..})(\bar{\mathbf{Y}}_{.j} - \bar{\mathbf{Y}}_{..})'.
\end{aligned}$$

Her er anvendt den sædvanlige notation

$$\begin{aligned}
\bar{\mathbf{Y}}_{..} &= \frac{1}{km}\sum_{i=1}^{k}\sum_{j=1}^{m}\mathbf{Y}_{ij} \\
\bar{\mathbf{Y}}_{i.} &= \frac{1}{m}\sum_{j=1}^{m}\mathbf{Y}_{ij}, \qquad i = 1,\ldots,k \\
\bar{\mathbf{Y}}_{.j} &= \frac{1}{k}\sum_{i=1}^{k}\mathbf{Y}_{ij}, \qquad j = 1,\ldots,m.
\end{aligned}$$

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

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

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

Der gælder nu

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

$$\{\mathbf{y}_{11},\ldots,\mathbf{y}_{km} \left| \frac{\det(\mathbf{q}_1)}{\det(\mathbf{q}_1+\mathbf{q}_2)} \leq U(p,k-1,(k-1)(m-1))_\alpha \right.\}.$$

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

$$\{\mathbf{y}_{11},\ldots,\mathbf{y}_{km} \left| \frac{\det(\mathbf{q}_1)}{\det(\mathbf{q}_1+\mathbf{q}_3)} \leq U(p,m-1,(k-1)(m-1))_\alpha \right.\}.$$

▲

BEVIS 6.14. Forbigås. Følger umiddelbart af sætning 6.10. Se f.eks. [3] ■

Vi samler resultaterne i et sædvanligt variansanalyseskema

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

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

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

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

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

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

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

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

dvs: vi kan ikke afvise, at β-erne er lig nul.
Tilsvarende er teststørrelsen for hypotesen $\alpha_1 = \alpha_2 = \alpha_3 = 0$ lig

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

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

$$\mathrm{F}_{97.5\%}(2,10) = 5.46,$$

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 ($y'_{ij} = 1000y_{ij}$)

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

$$\mathrm{F}_{99\%}(2,10) = 7,56.$$

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

$$\mathbf{Y}_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}, \qquad i = 1,2,3, \quad j = 1,\ldots,6,$$

hvor

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}, \quad \alpha_i = \begin{pmatrix} \alpha_{1i} \\ \alpha_{2i} \\ \alpha_{3i} \end{pmatrix}, \quad \beta_j = \begin{pmatrix} \beta_{1j} \\ \beta_{2j} \\ \beta_{3j} \end{pmatrix},$$

og observationerne er

$$\mathbf{Y}_{ij} = \begin{pmatrix} \text{grøntindhold} & \text{i plante } i \text{ på blok } j \\ \text{kvælstofindhold} & -''- \\ \text{tørstofindhold} & -''- \end{pmatrix}.$$

De realiserede udfald er

$$\mathbf{y}_{11} = \begin{pmatrix} 0.959 \\ 0.286 \\ 9.170 \end{pmatrix}, \ldots, \quad \mathbf{y}_{36} = \begin{pmatrix} 56.364 \\ 0.386 \\ 11.903 \end{pmatrix}.$$

Vi slår således de tre variansanalysemodeller ovenfor sammen i én.

Med notationen fra p. 297 findes de realiserede udfald af matricerne $\mathbf{Q}_1$, $\mathbf{Q}_2$ og $\mathbf{Q}_3$ til

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

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

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

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

Source	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 variationen mellem planter og J til variationen mellem blokke.

Den (her eksakte) F-teststørrelse for test af hypotesen $\alpha_1 = \alpha_2 = \alpha_3 = 0$, d.v.s. hypotesen, at alle planter er ens, er 43.6. Antallet af frihedsgrader er $(6, 16)$. Da

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

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

Da

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

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 planteydbytter. 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 $\mathbf{X} \in \mathrm{N}_p(\mu, \Sigma)$, og vi inddeler $\mathbf{X}$ i k komponenter med dimensionerne $p_1, \ldots, p_k$, d.v.s. :

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

Den tilsvarende spaltning af parametrene er

$$\mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_k \end{bmatrix}$$

og

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

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

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

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

SÆTNING 6.14. Vi betragter ovenstående situation og sætter

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

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

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

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

$$\begin{aligned} &P\{-m \log V \leq v\} \\ &\quad \simeq P\{\chi^2(f) \leq v\} + \frac{\gamma_2}{m^2}[P\{\chi^2(f+4) \leq v\} - P\{\chi^2(f) \leq v\}], \end{aligned}$$

hvor

$$\begin{aligned} m &= n - \frac{3}{2} - \frac{p^3 - \sum p_i^3}{3(p^2 - \sum p_i^2)} \\ \gamma_2 &= \frac{p^4 - \sum p_i^4}{48} - \frac{5(p^2 - \sum p_i^2)}{96} - \frac{(p^3 - \sum p_i^3)^2}{72(p^2 - \sum p_i^2)}. \end{aligned}$$

$$f = \frac{1}{2}[p^2 - \sum p_i^2], \qquad p = \sum p_i$$

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

BEVIS 6.15. Forbigås. Se f.eks. [3] ■

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 6.15. Vi betragter uafhængige observationer $\mathbf{X}_1, \ldots, \mathbf{X}_n$ med $\mathbf{X}_i \in \mathrm{N}_p(\mu, \boldsymbol{\Sigma})$, og vi sætter

$$\mathbf{A} = \sum (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})'.$$

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

$$\mathrm{W} = \frac{[\det(\mathbf{A}\,\boldsymbol{\Sigma}_0^{-1})]^{\frac{n}{2}}}{[\mathrm{tr}\,\mathbf{A}\,\boldsymbol{\Sigma}_0^{-1}/p]^{\frac{pn}{2}}}.$$

Ved fastlæggelse af det kritiske område kan benyttes, at

$$\begin{aligned} &P\{-(n-1)\rho \log \mathrm{W} \leq z\} \\ &\quad \simeq P\{\chi^2(f) \leq z\} + \omega_2[P\{\chi^2[f+4] \leq z\} - P\{\chi^2(f) \leq z\}], \end{aligned}$$

hvor

$$\begin{aligned} \rho &= 1 - \frac{2p^2 + p + 2}{6p(n-1)} \\ f &= \frac{1}{2}p(p+1) - 1 \\ \omega_2 &= \frac{(p+2)(p-1)(p-2)(2p^3 + 6p^2 + 3p + 2)}{288p^2n^2\rho^2}. \end{aligned}$$

▲

BEVIS 6.16. Forbigås. Se f.eks. [3]. ■

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

SÆTNING 6.16. Vi betragter indbyrdes uafhængige observationer $\mathbf{X}_1, \ldots, \mathbf{X}_n$ med $\mathbf{X}_i \in \mathrm{N}_p(\mu, \Sigma)$, og vi sætter

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

Kvotienttesstørrelsen for test af $H_0 : \Sigma = \Sigma_0$, hvor Σ_0 er kendt, mod alle alternativer, er

$$\lambda_1 = (\frac{e}{n})^{pn/2} [\det(\mathbf{A}\,\Sigma_0^{-1})]^{\frac{n}{2}} \exp(-\frac{1}{2}\operatorname{tr}(\mathbf{A}\,\Sigma_0^{-1})).$$

Ved fastlæggelse af det kritiske område kan benyttes, at

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

▲

BEVIS 6.17. Forbigås. Se f.eks. [3]. ■

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

$$\begin{array}{cccc} \mathbf{X}_{11}, & \ldots, & \mathbf{X}_{1n_1}, & \mathbf{X}_{1j} \in \mathrm{N}_p(\mu_1, \Sigma_1) \\ \vdots & & \vdots & \\ \mathbf{X}_{k1}, & \ldots, & \mathbf{X}_{kn_k}, & \mathbf{X}_{kj} \in \mathrm{N}_p(\mu_k, \Sigma_k) \end{array},$$

og vi ønsker at teste hypotesen

$$H_0 : \Sigma_1 = \cdots = \Sigma_k \quad \text{mod} \quad H_1 : \exists i, j : \Sigma_i \neq \Sigma_j.$$

Vi sætter

$$\begin{aligned} n &= \sum n_i, \\ \mathbf{A}_i &= \sum_{j=1}^{n_i} (\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)(\mathbf{X}_{ij} - \bar{\mathbf{X}}_i)', \end{aligned}$$

og

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

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

Vi har da

SÆTNING 6.17. Som teststørrelse for test af H_0 mod H_1 kan benyttes

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

Det kritiske område er af formen

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

og ved fastlæggelse af dette kan benyttes, at

$$P\{-2\rho \log W_1 \leq z\}$$
$$P\{\chi^2(f) \leq z\} + \omega_2[P\{\chi^2(f+4) \leq z\} - P\{\chi^2(f) \leq z\}],$$

hvor

$$\begin{aligned} f &= \frac{1}{2}(k-1)p(p+1), \\ \rho &= 1 - (\sum_i \frac{1}{n_i} - \frac{1}{n})\frac{2p^2+3p-1}{6(p+1)(k-1)}, \\ \omega_2 &= \frac{1}{48\rho^2}p(p+1)[(p-1)(p+2)(\sum_i \frac{1}{n_i^2} - \frac{1}{n^2}) - 6(k-1)(1-\rho)^2]. \end{aligned}$$

▲

BEVIS 6.18. Forbigås. Se f.eks. [3] ■

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

$$\mathbf{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}.$$

Hvis individet stammer fra π_1, er frekvensfunktionen for $\mathbf{X}$ $\mathrm{f}_1(\mathbf{x})$, og hvis det stammer fra π_2, er den $\mathrm{f}_2(\mathbf{x})$.

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 **beslutningsfunktionen** d: $R^p \to \{\pi_1, \pi_2\}$. d defineres ved

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

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

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

jvf. p. 6.6 i bind 1.

Det forventede tab i denne fordeling er

$$\begin{aligned} E_{\mathbf{x}}(L(\pi_i, d_{R_1}(\mathbf{x}))) &= L(\pi_1, d_{R_1}(\mathbf{x}))k(\pi_1|\mathbf{x}) + L(\pi_2, d_{R_1}(\mathbf{x}))k(\pi_2|\mathbf{x}) \\ &= \begin{cases} L(\pi_2, \pi_1)k(\pi_2|\mathbf{x}), & \mathbf{x} \in R_1 \\ L(\pi_1, \pi_2)k(\pi_1|\mathbf{x}), & \mathbf{x} \in R_2 \end{cases}. \end{aligned}$$

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

$$\begin{aligned} \mathbf{x} \in R_1 &\Leftrightarrow L(2,1)k(\pi_2|\mathbf{x}) \leq L(1,2)k(\pi_1|\mathbf{x}) \\ &\Leftrightarrow \frac{L(1,2)f_1(\mathbf{x})p_1}{L(2,1)f_2(\mathbf{x})p_2} \geq 1 \\ &\Leftrightarrow \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{L(2,1)}{L(1,2)}\frac{p_2}{p_1}. \end{aligned}$$

Vi samler disse overvejelser i

SÆTNING 7.1. Bayesløsningen til klassifikationsproblemet er givet ved området

$$R_1 = \{\mathbf{x}|\frac{\mathrm{f}_1(\mathbf{x})}{\mathrm{f}_2(\mathbf{x})} \geq \frac{\mathrm{L}(2,1)}{\mathrm{L}(1,2)}\frac{p_2}{p_1}\}.$$

▲

BEMÆRKNING 7.1. 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)

$$\begin{aligned} R(\pi_1, \mathrm{d}_{R_1}) &= \mathrm{E}_{\pi_1}\mathrm{L}(\pi_1, \mathrm{d}_{R_1}(\mathbf{X})) = \mathrm{L}(1,2)P\{\mathbf{X} \in R_2|\pi_1\}. \\ R(\pi_2, \mathrm{d}_{R_1}) &= \mathrm{E}_{\pi_2}\mathrm{L}(\pi_2, \mathrm{d}_{R_1}(\mathbf{X})) = \mathrm{L}(2,1)P\{\mathbf{X} \in R_1|\pi_2\}. \end{aligned}$$

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

$$R_1 = \{\mathbf{x}|\frac{\mathrm{f}_1(\mathbf{x})}{\mathrm{f}_2(\mathbf{x})} \geq c\},$$

hvor c bestemmes ved

$$\mathrm{L}(1,2)P\{\frac{\mathrm{f}_1(\mathbf{x})}{\mathrm{f}_2(\mathbf{x})} < c|\pi_1\} = \mathrm{L}(2,1)P\{\frac{\mathrm{f}_1(\mathbf{X})}{\mathrm{f}_2(\mathbf{X})} \geq c|\pi_2\}.$$

▲

BEMÆRKNING 7.2. Relationen til bestemmelse af c kan skrives

$$\begin{aligned} & \mathrm{L}(1,2) \cdot \text{(sandsynligheden for misklassifikation hvis } \pi_1\text{er sand)} \\ = \ & \mathrm{L}(2,1) \cdot \text{(sandsynligheden for misklassifikation hvis } \pi_2\text{er sand)} \end{aligned}$$

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 betrager 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 $\pi_1 \simeq N(\mu_1, \Sigma)$ og $\pi_2 \simeq N(\mu_2, \Sigma)$. Da gælder

$$\frac{f_1(x)}{f_2(x)} \geq c \Leftrightarrow x'\Sigma^{-1}(\mu_1 - \mu_2) - \frac{1}{2}\mu_1'\Sigma^{-1}\mu_1 + \frac{1}{2}\mu_2'\Sigma^{-1}\mu_2 \geq \log c.$$

▲

BEVIS 7.1. Vi indfører det indre produkt $(\cdot|\cdot)$ og normen $\| \quad \|$ ved

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

og

$$\|x\|^2 = (x|x).$$

Vi har da

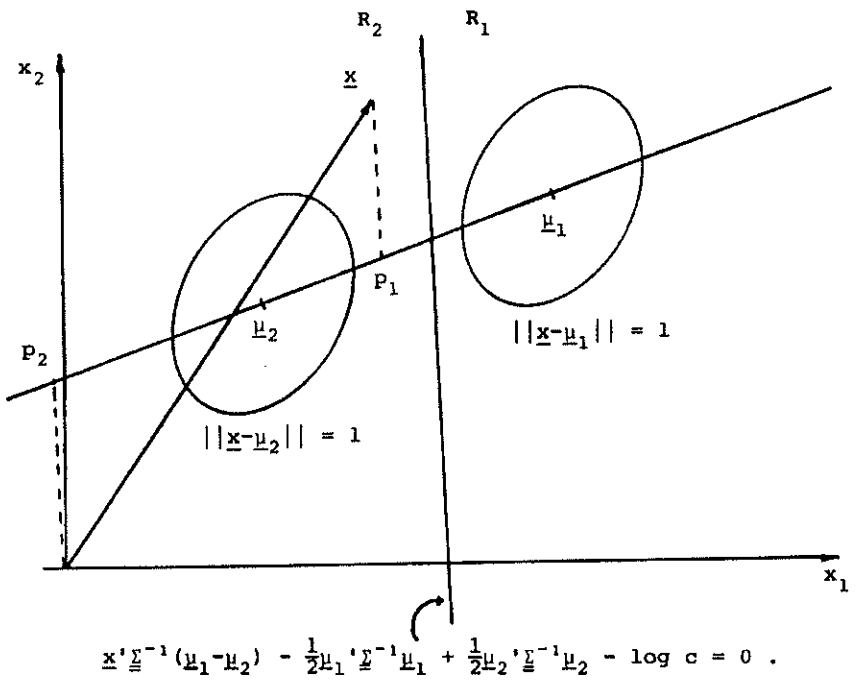
$$f_i(x) = \frac{1}{\sqrt{2\pi}^p \sqrt{\det \Sigma}} \exp(-\frac{1}{2}\|x - \mu_i\|^2).$$

Heraf fås umiddelbart

$$\begin{aligned}
& \frac{f_1(x)}{f_2(x)} \geq c \Leftrightarrow \log \frac{f_1(x)}{f_2(x)} \geq \log c \\
\Leftrightarrow \quad & -\|x - \mu_1\|^2 + \|x - \mu_2\|^2 \geq 2\log c \\
\Leftrightarrow \quad & -(x - \mu_1|x - \mu_1) + (x - \mu_2|x - \mu_2) \geq 2\log c \\
\Leftrightarrow \quad & 2(x|\mu_1) - 2(x|\mu_2) - (\mu_1|\mu_1) + (\mu_2|\mu_2) \geq 2\log c \\
\Leftrightarrow \quad & 2(x|\mu_1 - \mu_2) - (\mu_1|\mu_1) + (\mu_2|\mu_2) \geq 2\log c.
\end{aligned}$$

Ved at benytte sammenhængen mellem $(|)$ og Σ^{-1} fås sætningen umiddelbart. ■

BEMÆRKNING 7.3. Udtrykket $\frac{f_1(x)}{f_2(x)} \geq c$ 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)



Vektoren $p_1\vec{p}_2$ betegner den ortogonale projektion (NB! ortogonal med hensyn til Σ^{-1}) af x på linien, der forbinder μ_1 og μ_2. (Det kan vises, at hældningen af projektionslinierne m.v. er lig hældningen af ellipse (ellipsoide) tangenterne i de punkter, hvor de skærer linien (μ_1, μ_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 x er tilstrækkelig lang (regnet med fortegn). Ellers klassificerer vi observationen som stammende fra π_2.

Funktionen

$$\mathbf{x}'\Sigma^{-1}(\mu_1 - \mu_2) - \frac{1}{2}\mu_1'\Sigma^{-1}\mu_1 + \frac{1}{2}\mu_2'\Sigma^{-1}\mu_2 - \log c$$

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

$$\delta = \Sigma^{-1}(\mu_1 - \mu_2),$$

har vi følgende

SÆTNING 7.4. Vektoren δ har den egenskab, at den maksimaliserer funktionen

$$\varphi(\mathbf{d}) = \frac{[\mathrm{E}_1(\mathbf{X}'\mathbf{d}) - \mathrm{E}_2(\mathbf{X}'\mathbf{d})]^2}{\mathrm{V}(\mathbf{X}'\mathbf{d})} = \frac{[(\mu_1 - \mu_2)'\mathbf{d}]^2}{\mathbf{d}'\Sigma\mathbf{d}}.$$

▲

BEVIS 7.2. Beviset er ikke særlig interessant, men relativt simpelt. Da vi umiddelbart ser, at $\varphi(k \cdot d) = k \cdot \varphi(\mathbf{d})$, kan vi bestemme ekstremaer for φ ved at bestemme ekstremaer for tælleren under bibetingelsen

$$\mathbf{d}'\Sigma\mathbf{d} = 1.$$

Vi indfører en Lagrange multiplikator λ og søger maksimum af

$$\psi(\mathbf{d}) = [(\mu_1 - \mu_2)'d]^2 - \lambda(\mathbf{d}'\Sigma\mathbf{d} - 1).$$

Nu er

$$\frac{\partial\psi}{\partial\mathbf{d}} = 2(\mu_1 - \mu_2)(\mu_1 - \mu_2)'\mathbf{d} - 2\lambda\Sigma\mathbf{d}.$$

Sættes denne lig 0, fås

$$(\mu_1 - \mu_2)(\mu_1 - \mu_2)'\mathbf{d} = \lambda\Sigma\mathbf{d},$$

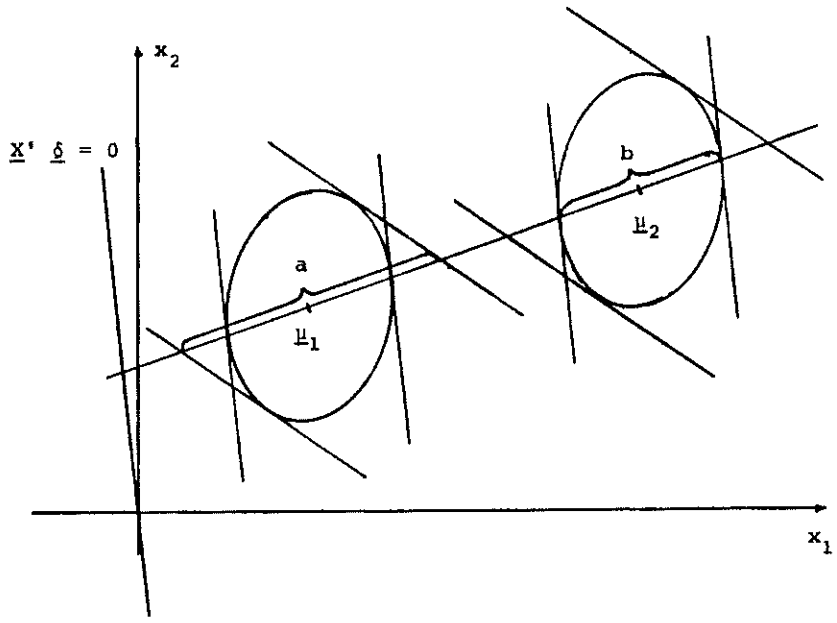
d.v.s.

$$\mathbf{d} = \frac{(\mu_1 - \mu_2)'\mathbf{d}}{\lambda}\Sigma^{-1}(\mu_1 - \mu_2) = k \cdot \delta,$$

hvor k er en skalar. ■

BEMÆRKNING 7.4. Sætningens indhold er, at den ved δ bestemte lineære funktion

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



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 μ_1, μ_2 efter retningen bestemt ved $\mathbf{x}'\delta = 0$

a: projektionen af ellipsen på linien μ_1, μ_2 efter anden retning.

Det fremgår, at den ved δ bestemte projektion på linien, der forbinder μ_1 og μ_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

$$Z = \mathbf{X}'\Sigma^{-1}(\mu_1 - \mu_2) - \frac{1}{2}\mu_1'\Sigma^{-1}\mu_1 + \frac{1}{2}\mu_2'\Sigma^{-1}\mu_2.$$

Om dette gælder

$$Z \in \begin{cases} \mathrm{N}(+\frac{1}{2}\|\mu_1-\mu_2\|^2, \|\mu_1-\mu_2\|^2), & \text{hvis } \pi_1 \text{ sand} \\ \mathrm{N}(-\frac{1}{2}\|\mu_1-\mu_2\|^2, \|\mu_1-\mu_2\|^2), & \text{hvis } \pi_2 \text{ sand} \end{cases}.$$

▲

BEVIS 7.3. Beviset er helt ligefremt. Lad os e.g. betragte tilfældet π_1 sand. Vi har da $\mathrm{E}(\mathbf{X}) = \mu_1$ og dermed

$$\begin{aligned} \mathrm{E}(Z) &= \mu_1'\Sigma^{-1}(\mu_1-\mu_2) - \frac{1}{2}\mu_1'\Sigma^{-1}\mu_1 + \frac{1}{2}\mu_2'\Sigma^{-1}\mu_2 \\ &= \frac{1}{2}(\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2) \\ &= \frac{1}{2}\|\mu_1-\mu_2\|^2. \end{aligned}$$

$$\begin{aligned} \mathrm{V}(Z) &= (\mu_1-\mu_2)'\Sigma^{-1}\Sigma\Sigma^{-1}(\mu_1-\mu_2) \\ &= (\mu_1-\mu_2)'\Sigma^{-1}(\mu_1-\mu_2) \\ &= \|\mu_1-\mu_2\|^2. \end{aligned}$$

Resultatet vedrørende π_2 vises ganske analogt. ■

Vi ser nu på en række eksempler.

EKSEMPEL 7.1. Vi betragter det tilfælde, hvor

$$\begin{aligned} \pi_1 &\leftrightarrow \mathrm{N}(\begin{pmatrix} 4 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}) \\ \pi_2 &\leftrightarrow \mathrm{N}(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}), \end{aligned}$$

og vi vil bestemme en "bedste" diskriminatorfunktion. Da der intet er oplyst om a priori sandsynligheder og lignede, 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

$$(x_1 x_2)\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 3 \\ 1 \end{pmatrix} - \frac{1}{2}(2\cdot 16 + 1\cdot 4 - 2\cdot 8) + \frac{1}{2}(2\cdot 1 + 1\cdot 1 - 2\cdot 1) = 0$$

eller

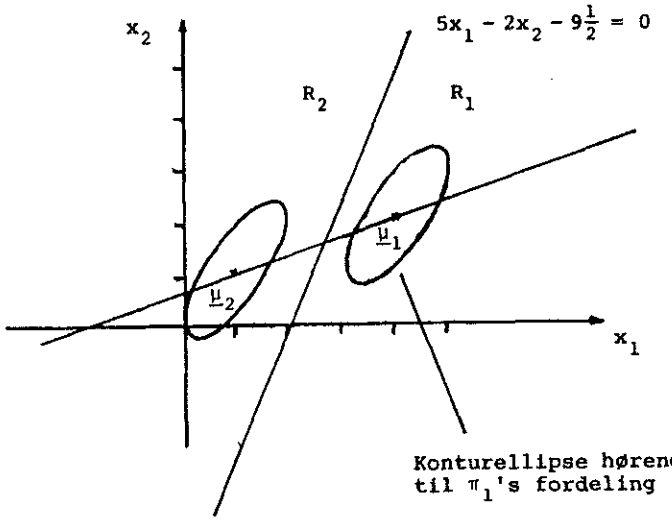
$$5x_1 - 2x_2 - 9\frac{1}{2} = 0.$$

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

$$5\cdot 5 - 2\cdot 6 - 9\frac{1}{2} = 3\frac{1}{2} > 0.$$

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

$$\|\mu_1 - \mu_2\|^2 = 2 \cdot 9 + 1 \cdot 1 - 2 \cdot 3 \cdot 1 = 13.$$

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

$$\begin{aligned}
& 2 \cdot P\left\{\frac{\mathrm{f}_1(\mathbf{X})}{\mathrm{f}_2(\mathbf{X})} < c|\pi_1\right\} = P\left\{\frac{\mathrm{f}_1(\mathbf{X})}{\mathrm{f}_2(\mathbf{X})} \geq c|\pi_2\right\} \\
\Leftrightarrow \quad & 2 \cdot P\{Z < \log c|\pi_1\} = P\{Z \geq \log c|\pi_2\} \\
\Leftrightarrow \quad & 2 \cdot P\{\mathrm{N}(\tfrac{1}{2}13, 13) < \log c\} = P\{\mathrm{N}(-\tfrac{1}{2}13, 13) \geq \log c\} \\
\Leftrightarrow \quad & 2 \cdot P\left\{\mathrm{N}(0,1) < \frac{\log c - 6.5}{\sqrt{13}}\right\} = P\left\{\mathrm{N}(0,1) \geq \frac{\log c + 6.5}{\sqrt{13}}\right\}.
\end{aligned}$$

Ved at prøve med forskellige værdier af c indses, at

$$c \simeq 0.5617.$$

Med denne værdi er misklassifikationssandsynlighederne

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

$$\text{Hvis } \pi_2 \text{ sand:} \quad P\{\mathrm{N}(0,1) < \tfrac{\log 0.5617 + 6.5}{\sqrt{13}}\} \simeq 0.050.$$

Den diskriminerende linie bliver nu bestemt ved

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

eller

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

Denne linie skærer forbindelseslinien mellem μ_1 og μ_2 i $(2.36, 1.46)$, d.v.s. rykket mod μ_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 μ_1 er sand, end hvis μ_2 er det. Vi skal derfor gøre R_1 større, i.e. rykke den begrænsende linie mod μ_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.

$$\pi_1 \leftrightarrow \mathrm{N}\left(\begin{pmatrix} 4 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}\right)$$
$$\pi_2 \leftrightarrow \mathrm{N}\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$$

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

For $c > 0$ har vi

$$\frac{\mathrm{f}_1(\mathbf{x})}{\mathrm{f}_2(\mathbf{x})} \geq c \quad \Leftrightarrow$$

$$-(\mathbf{x}-\mu_1)'\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\mu_1) + (\mathbf{x}-\mu_2)'\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\mu_2) \geq 2\log c.$$

Da

$$\begin{aligned}(\mathbf{x}-\mu_1)'\boldsymbol{\Sigma}_1^{-1}(\mathbf{x}-\mu_1) &= 2(x_1-4)^2 - (x_2-2)^2 - 2(x_1-4)(x_2-2) \\ &= 2x_1^2 + x_2^2 - 2x_1x_2 - 12x_1 + 4x_2 + 20,\end{aligned}$$

og

$$\begin{aligned}(\mathbf{x}-\mu_2)'\boldsymbol{\Sigma}_2^{-1}(\mathbf{x}-\mu_2) &= (x_1-1)^2 + (x_2-1)^2 \\ &= x_1^2 + x_2^2 - 2x_1 - 2x_2 + 2,\end{aligned}$$

er

$$\frac{\mathrm{f}_1(\mathbf{x})}{\mathrm{f}_2(\mathbf{x})} \geq c \Leftrightarrow -x^2 + 2x_1x_2 + 10x_1 - 6x_2 - 18 \geq 2\log c.$$

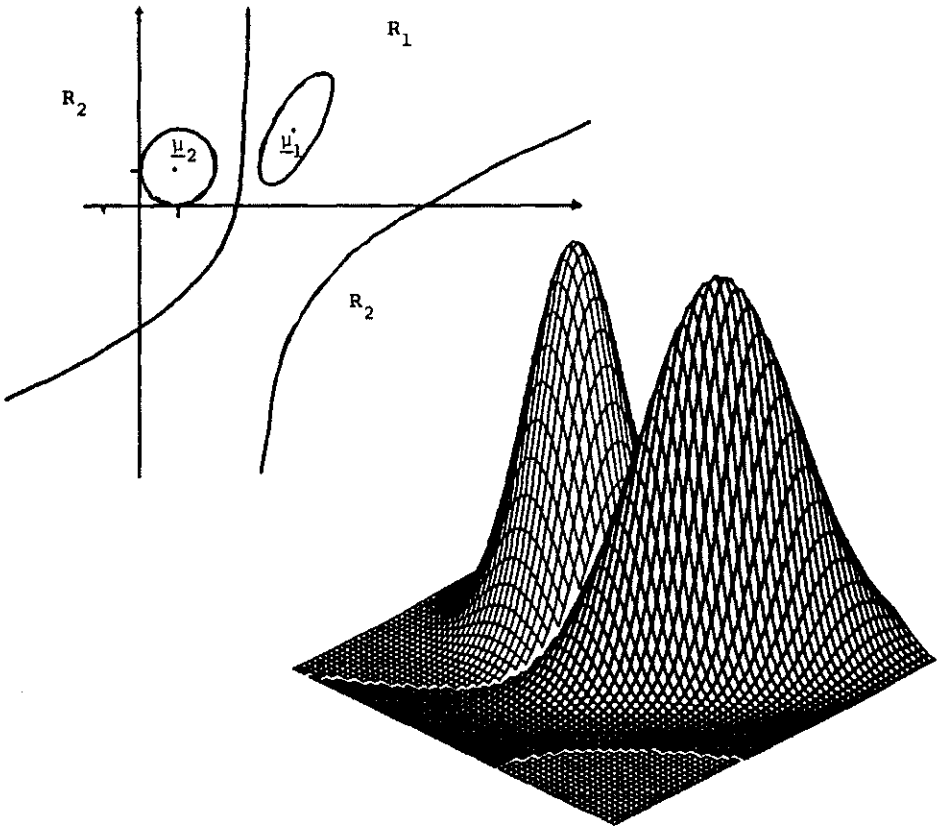
Vælger vi $c = 1$, ser vi, at kurven, der adskiller R_1 og R_2, er hyperblen

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

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 eksakte.

Lad os betragte det normale tilfælde

$$\begin{aligned} \pi_1 &\leftrightarrow \mathrm{N}(\mu_1, \Sigma) \\ \pi_2 &\leftrightarrow \mathrm{N}(\mu_2, \Sigma), \end{aligned}$$

hvor parametrene er ukendte. Hvis vi har observationer $\mathbf{X}_1, \ldots, \mathbf{X}_{n_1}$, som vi **ved** stammer fra π_1, og observationer $\mathbf{Y}_1, \ldots, \mathbf{Y}_{n_2}$, som vi **ved** stammer fra π_2, kan vi estimere parametrene som følger

$$\begin{aligned} \hat{\mu}_1 &= \frac{1}{n_1}\sum_i \mathbf{X}_i = \bar{\mathbf{X}} \\ \hat{\mu}_2 &= \frac{1}{n_2}\sum_i \mathbf{Y}_i = \bar{\mathbf{Y}} \\ \hat{\Sigma} &= \frac{1}{n_1+n_2-2}\Big(\sum_i (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})' + \sum_i (\mathbf{Y}_i - \bar{\mathbf{Y}})(\mathbf{Y}_i - \bar{\mathbf{Y}})'\Big) \end{aligned}$$

Vi har nu i fuldstændig analogi med sætningen p. 312 diskriminatoren

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

Den eksakte fordeling af denne størrelse, hvis vi erstatter $\mathbf{x}$ med en stokastisk variabel $\mathbf{X} \in \mathrm{N}(\mu_i, \Sigma)$, 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

$$\|\hat{\mu}_1 - \hat{\mu}_2\|^2 \simeq \mathrm{D}^2 = (\hat{\mu}_1 - \hat{\mu}_2)'\hat{\Sigma}^{-1}(\hat{\mu}_1 - \hat{\mu}_2) = \|\hat{\mu}_1 - \hat{\mu}_2\|^2_{\hat{\Sigma}^{-1}}.$$

kaldes **Mahalanobis'** afstand. Det må her indskydes, at en mængde forfattere benytter vendingen Mahalanobis' afstand også om størrelsen $\|\mu_1 - \mu_2\|^2$ 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 $\mu_1 = \mu_2$, idet

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

er $\mathrm{F}(p, n_1 + n_2 - p - 1)$-fordelt, hvis $\mu_1 = \mu_2$. Hvis $\mu_1 \neq \mu_2$, 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: matsimal bredde af hovedet iden genale region
x_3: længd 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 $\mu_1 = \mu_2$. Vi har

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

Denne værdi indsætter vi i teststørrelsen p. 321 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

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

og

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

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

$$\mathrm{d}(\mathbf{x}) = -2.7458x_1 + 6.6217x_2 + 4.582x_3 - 25.3506$$

ved klassifikation af de to mulige græshoppearter.

Har vi eksempelvis fanget et eksemplar med de målte karakteristika

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

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

7.1.4 Test for bedste diskriminantfunktion

Vi minder om, at den bedste diskriminator

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

kan fås ved maksimalisering af funktionen

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

Maksimalværdien er

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

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

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

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

$$Z = \frac{n_1 + n_2 - p - 1}{p - 1} \cdot \frac{n_1 n_2(\mathrm{D}^2 - \mathrm{D}_1^2)}{(n_1 + n_2)(n_1 + n_2 - 2) + n_1 n_2\,\mathrm{D}_1^2},$$

der er $\mathrm{F}(p-1, n_1+n_2-p-1)$-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 $\mathbf{d}$ i stedet for $\hat{\delta}$. Hvis denne formindskelse er for stor, i.e. hvis Z er stor, vil vi ikke kunnne antage, at $\mathbf{d}$ yder en lige så god skelnen mellem de to populationer som $\hat{\delta}$.

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

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

Heraf findes umiddelbart

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

Mahalanobis' afstand mellem middelværdierne er

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

Vi tester først, om det kan antages, at $\mu_1 = \mu_2$. Teststørrelsen bliver

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

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

Det vil ikke være rimeligt at antage $\mu_1 = \mu_2$.

Ved at se på differenserne mellem komponenterne i μ_1 og μ_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 $\mu_1 - \mu_2$, kunne man prøve med afbildningen

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

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

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

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

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

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

Teststørrelsen bliver

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

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

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 virkeligt har været og er nødvendigt med alle målinger, eller om man kan nøjes med færre variable til at skille populationer fra hinanden. Man kunne eksempelvis forestille sig, at det ville være tilstrækkeligt at måle **længden** 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 $X_1, \ldots, X_p$. 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 populationen π_1 og n_2 observationer fra populationen π_2. Vi sætter

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

og foretager samme opspaltning af middelværdivektor og dispersionsmatrix

$$\mu_i = \begin{bmatrix} \mu_i^{(1)} \\ \mu_i^{(2)} \end{bmatrix}$$

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

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

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

og

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

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 $Z \in \mathrm{F}(q, n_1 + n_2 - p - 1)$, 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

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

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

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**

$$\pi_1, \ldots, \pi_k$$

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

Måleresultatet er

$$\mathbf{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_p \end{pmatrix}.$$

Hvis individet stammer fra π_i, er frekvensfunktionenen for $\mathbf{X}$ $\mathrm{f}_i(\mathbf{x})$.

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	$\mathrm{L}(1,2)$	$\cdots$	$\mathrm{L}(1,k)$
	π_2	$\mathrm{L}(2,1)$	0	$\cdots$	$\mathrm{L}(2,k)$
	$\vdots$	$\vdots$	$\vdots$		$\vdots$
	π_k	$\mathrm{L}(k,1)$	$\mathrm{L}(k,2)$	$\cdots$	0

Endelig kan vi antage, at der er en **a priorifordeling**

$$\mathrm{g}(\pi_i) = p_i, \qquad i = 1, \ldots, k.$$

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

$$S_i^*(\mathbf{x}) = S_i^* = -[p_1\mathrm{f}_1(\mathbf{x})\mathrm{L}(1,i) + \cdots + p_k\mathrm{f}_k(\mathbf{x})\mathrm{L}(k,i)]$$

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

$$\begin{aligned} \mathrm{k}(\pi_\nu|\mathbf{x}) &= \frac{p_\nu \mathrm{f}_\nu \mathbf{x})}{p_1\mathrm{f}_1(\mathbf{x}) + \cdots + p_k\mathrm{f}_k(\mathbf{x})} \\ &= \frac{p_\nu \mathrm{f}_\nu(\mathbf{x})}{\mathrm{h}(\mathbf{x})}, \end{aligned}$$

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

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

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

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

d.v.s. hvis

$$\begin{aligned} & -(\sum_\nu p_\nu \mathrm{f}_\nu(\mathbf{x}) - p_i \mathrm{f}_i(\mathbf{x})) > -(\sum_\nu p_\nu \mathrm{f}_\nu(\mathbf{x}) - p_j \mathrm{f}_j(\mathbf{x})) \\ \Leftrightarrow \quad & p_i \mathrm{f}_i(\mathbf{x}) > p_j \mathrm{f}_j(\mathbf{x}). \end{aligned}$$

Vi kan derfor i dette tilfælde vælge diskriminantværdien

$$S_i' = p_i \mathrm{f}_i(\mathbf{x}).$$

Bayesreglen er altså her, at vi vælger den population, der har den største a posteriori-sandsynlighed, dvs: 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

$$S_i'' = \mathrm{f}_i(\mathbf{x}),$$

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

$$\pi_i \quad \leftrightarrow \quad \mathrm{N}(\mu_i, \Sigma_i),$$

d.v.s.

$$\mathrm{f}_i(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^p}\frac{1}{\sqrt{\det \boldsymbol{\Sigma}_i}}\exp(-\frac{1}{2}(\mathbf{x}-\mu_i)'\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\mu_i)),$$

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

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 $(2\pi)^{-\frac{p}{2}}$. Dette giver (idet vi forudsætter, at tabene er ens)

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

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

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

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

$$S_i = \mathbf{x}'\boldsymbol{\Sigma}^{-1}\mu_i - \frac{1}{2}\mu_i'\boldsymbol{\Sigma}^{-1}\mu_i + \log p_i.$$

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

$$\begin{aligned} & S_1' > S_2' \Leftrightarrow S_1 - S_2 > 0 \\ \Leftrightarrow\quad & \mathbf{x}'\boldsymbol{\Sigma}^{-1}(\mu_1-\mu_2) - \frac{1}{2}\mu_1'\boldsymbol{\Sigma}^{-1}\mu_1 + \frac{1}{2}\mu_2'\boldsymbol{\Sigma}^{-1}\mu_2 \geq \log\frac{p_2}{p_1}, \end{aligned}$$

d.v.s. det samme resultat som p. 312.

A posteriori-sandsynligheden for den ν'te gruppe bliver

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

Det er naturligvis muligt at beskrive beslutningsreglerne ved en inddeling af R^p i mængder $R_1, \ldots, R_k$, således at vi vælger π_i, netop når $\mathbf{x} \in R_i$. 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

$$\mu_1 = \begin{pmatrix} 4 \\ 2 \end{pmatrix}, \quad \mu_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \text{og} \quad \mu_3 = \begin{pmatrix} 2 \\ 6 \end{pmatrix},$$

og den fælles dispersionsmatrix

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

jvf. eksemplet p. 316. Vi har da - idet vi antager, at alle p_i er ens, hvorfor vi kan udelade dem fra diskriminantværdierne -

$$\begin{aligned} S'_{11} &= (x_1 x_2)\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 4 \\ 2 \end{pmatrix} - \frac{1}{2}(4,2)\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 4 \\ 2 \end{pmatrix} \\ &= 6x_1 - 2x_2 - 10 \end{aligned}$$

$$\begin{aligned} S'_{12} &= (x_1 x_2)\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} - \frac{1}{2}(1,1)\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ &= x_1 - \frac{1}{2} \end{aligned}$$

$$\begin{aligned} S'_{13} &= (x_1 x_2)\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 2 \\ 6 \end{pmatrix} - \frac{1}{2}(2,6)\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 2 \\ 6 \end{pmatrix} \\ &= -2x_1 + 4x_2 - 10. \end{aligned}$$

Vi vælger nu π_1 frem for π_2, hvis

$$\begin{aligned} u_{12}(\mathbf{x}) &= 6x_1 - 2x_2 - 10 - (x_1 - \frac{1}{2}) \\ &= 5x_1 - 2x_2 - 9\frac{1}{2} \\ &> 0. \end{aligned}$$

Vi vælger π_1 frem for π_3, hvis

$$\begin{aligned} u_{13}(\mathbf{x}) &= 6x_1 - 2x_2 - 10 - (-2x_1 + 4x_2 - 10) \\ &= 8x_1 - 6x_2 \\ &> 0, \end{aligned}$$

og endelig π_2 frem for π_3, hvis

$$\begin{aligned} u_{23}(\mathbf{x}) &= x_1 - \frac{1}{2} - (-2x_1 + 4x_2 - 10) \\ &= 3x_1 - 4x_2 + 9\frac{1}{2} \\ &> 0. \end{aligned}$$

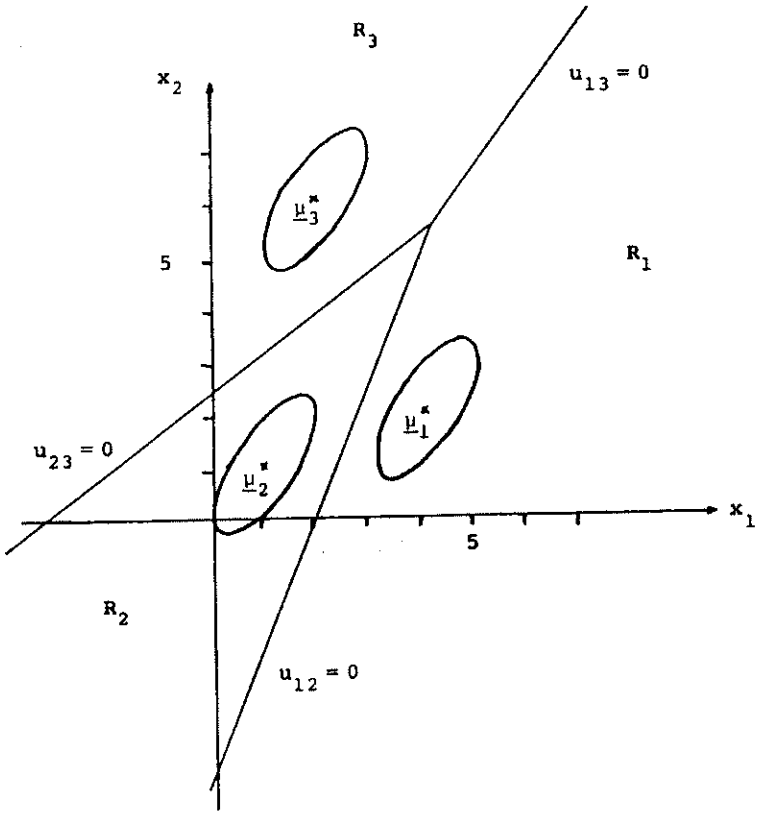
Det er nu evident, at vi vælger π_1, hvis såvel $u_{12}(\mathbf{x}) > 0$ som $u_{13}(\mathbf{x}) > 0$, og analogt med de øvrige.

Vi kan derfor definere områderne

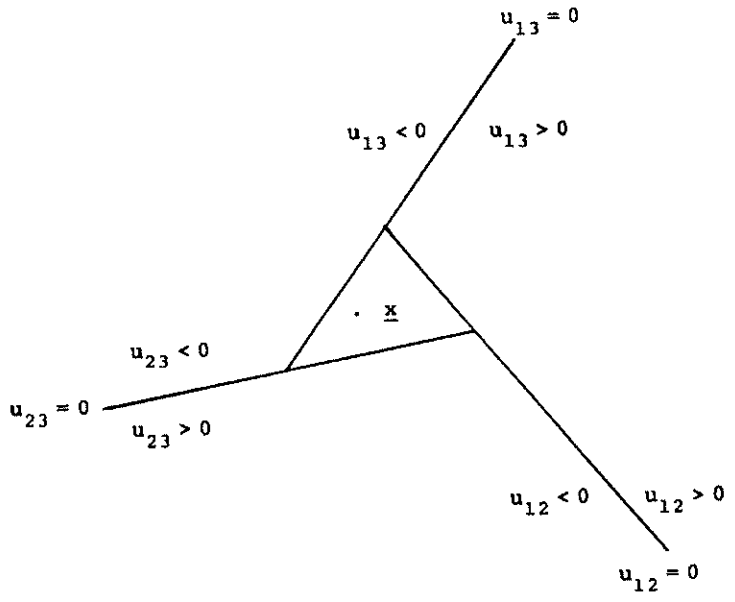
$$\begin{aligned} R_1 &= \{\mathbf{x} | u_{12}(\mathbf{x}) > 0 \quad \wedge \quad u_{13}(\mathbf{x}) > 0\} \\ R_2 &= \{\mathbf{x} | u_{12}(\mathbf{x}) < 0 \quad \wedge \quad u_{23}(\mathbf{x}) > 0\} \\ R_3 &= \{\mathbf{x} | u_{13}(\mathbf{x}) < 0 \quad \wedge \quad u_{23}(\mathbf{x}) < 0\}, \end{aligned}$$

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

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 skitseret på figur 7.1.



Figur 7.1:

Vi bemærker nu, at

$$u_{ij} > 0 \Leftrightarrow S'_{1i} > S'_{1j} \Leftrightarrow \mathrm{f}_i > \mathrm{f}_j .$$

Om punktet x gælder

$$\left.\begin{array}{lll} u_{23}(\mathbf{x}) < 0 & \text{d.v.s.} & \mathrm{f}_2(\mathbf{x}) < \mathrm{f}_3(\mathbf{x}) \\ u_{13}(\mathbf{x}) > 0 & \text{d.v.s.} & \mathrm{f}_1(\mathbf{x}) > \mathrm{f}_3(\mathbf{x}) \end{array}\right\} \Rightarrow \mathrm{f}_1(\mathbf{x}) > \mathrm{f}_2(\mathbf{x})$$
$$u_{12}(\mathbf{x}) < 0 \quad \text{d.v.s.} \quad \mathrm{f}_1(\mathbf{x}) < \mathrm{f}_2(\mathbf{x})$$

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 $n_1, \ldots, n_k$ observationer i hver. Gruppegennemsnittene kaldes $\bar{\mathbf{X}}_1, \ldots, \bar{\mathbf{X}}_k$. Vi definerer en "mellem grupper" matrix (among groups)

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

en "inden for grupper" matrix (within groups)

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

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

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

En fundamental ligning er

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

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 $\mathbf{y} = \mathbf{d}'\mathbf{x}$, så

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

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

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

eller

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

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

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

maksimaliseres under bibetingelsen, at

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

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

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

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

Projektionernes koordinater bliver

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

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 $(\mathbf{d}_1, \mathbf{d}_2)$-planen.

I f.eks. programmet BMD07M - STEPWISE DISCRIMINANT ANALYSIS - benævnes $(\mathbf{d}_1, \mathbf{d}_2)$ 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

$$H_0 : \mu_1 = \cdots = \mu_k \quad \text{mod} \quad H_1 : \exists i, j : \mu_i \neq \mu_j,$$

er

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

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.

EKSEMPEL 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.

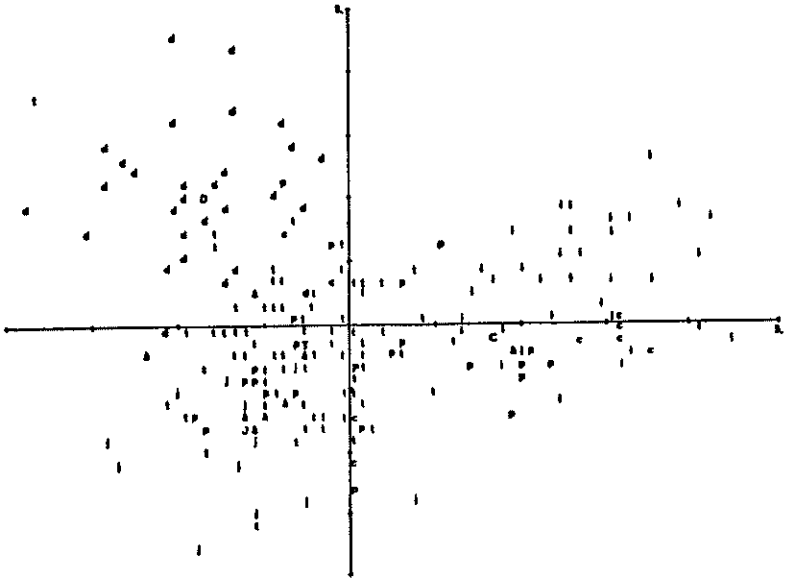
Variable	Middel	Stand.afv.
B	73	141
Ti	40563	22279
V	678	491
Cr	1135	1216
Mn	2562	2081
Fe	225817	122302
Co	62	26
Ni	116	54
Cu	69	56
Ga	21	10
Zr	14752	14771
Mo	29	20
Sn	56	99
Pb	351	786
Sum	–	–

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.

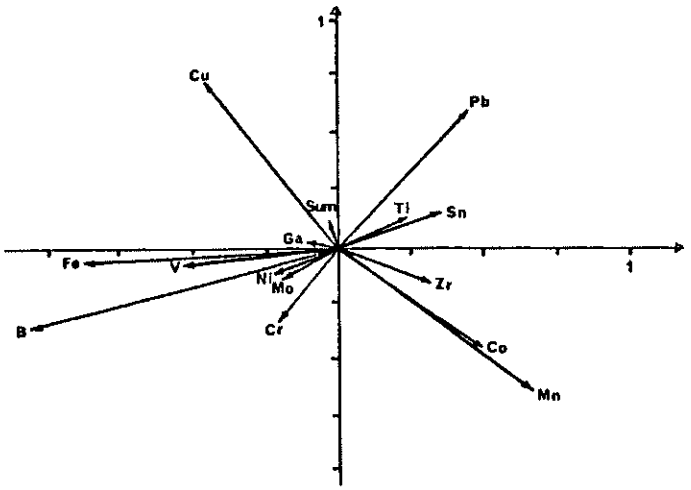
Periode	Antal
Jura	17
Trias	80
Perm	30
Kul	9
Devon	31
Tertiære intrusiver	35
Caledonsk krystallinsk	4
Eleonora Bay Formation	2

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, se figur 7.2.

I figur 7.3 er vist koefficienterne for de almindelige variable på de to "kanoniske" variable.



Figur 7.2:



Figur 7.3:

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 (BMD07M).

SSP programmet er opbygget på samme måde som BMD05M, 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 input-materiale.

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

Lad der være givet k populationer $\pi_1, \ldots, \pi_k$, og lad $\pi_i \leftrightarrow \mathrm{N}_p(\mu_i, \Sigma)$. Vi forudsætter, at der foreligger observationer

$$\begin{array}{lll} \mathbf{X}_{11} & \cdots & \mathbf{X}_{1n_1} \quad \text{fra } \pi_1 \\ \vdots & & \vdots \\ \mathbf{X}_{k1} & \cdots & \mathbf{X}_{kn_k} \quad \text{fra } \pi_k. \end{array}$$

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

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

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

Vi vil nu bestemme den approxinative fordeling for V. Hvis $\mathbf{X} \in \mathrm{N}_p(\mu, \Sigma)$, gælder ifølge sætning 2.15, at

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

Af χ^2-fordelingens reproduktivitetssætning fås - såfremt alle $\mu_i = \mu$ -

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

og dermed, at

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

Estimeres nu de p værdier i μ, fås stadig en approximativ χ^2-fordeling, men med p frihedsgrader færre, i.e.

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

såfremt alle μ_i er ens. **Størrelsen V kan derfor bruges som teststørrelse** ved test af hypotesen

$$H_0 : \mu_1 = \cdots = \mu_k \quad \text{mod} \quad H_1 : \exists i, j(\mu_i \neq \mu_j).$$

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

$$\begin{aligned} V_2 &= \sum_{i=1}^{2} n_i(\bar{\mathbf{X}}_i - \bar{\mathbf{X}})'\hat{\Sigma}^{-1}(\bar{\mathbf{X}}_i - \bar{\mathbf{X}}) \\ &= \frac{n_1 n_2}{n_1 + n_2}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)'\hat{\Sigma}^{-1}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2), \end{aligned}$$

eller

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

Kalder vi den i'te diskriminantfunktion (jvf. p. 330) g_i, d.v.s.

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

kan man for hver observation beregne størrelsen

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

hvor

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

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

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

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

Hvis p_{ij} er nær 0, der

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

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

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.

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

En udskrift er vist p. 340 og 340.

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 udføldes som følger

skema

Hvis vi benævner observationerne

observationer

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, F4.0, 4x)

Vi vil nu vise et par eksempler på kørsel med MDISC

EKSEMPEL 7.9. Vi betragter data fra den p. 324 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.

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 $\mu_1 = \mu_2 = \mu_3$ 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. 341. ♦

EKSEMPEL 7.10. Vi betragter nu problemet, der er anført i eksemplet p. 327, 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. Derneæst ændres statement 5 tilbage til det oprindelige:

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

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. 327. (Den der anførte afstand er 76.7082. Afvigelsen skyldes afrundingsfejl).

♦

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 århundredeskiftet (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

$$\mathbf{X} = \begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix},$$

der har dispersionsmatricen

$$\mathrm{D}(X) = \mathbf{\Sigma},$$

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

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

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

De tilsvarende ortonormerede egenvektorer benævnes

$$\mathbf{p}_1, \ldots, \mathbf{p}_k,$$

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

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

Vi har da følgende

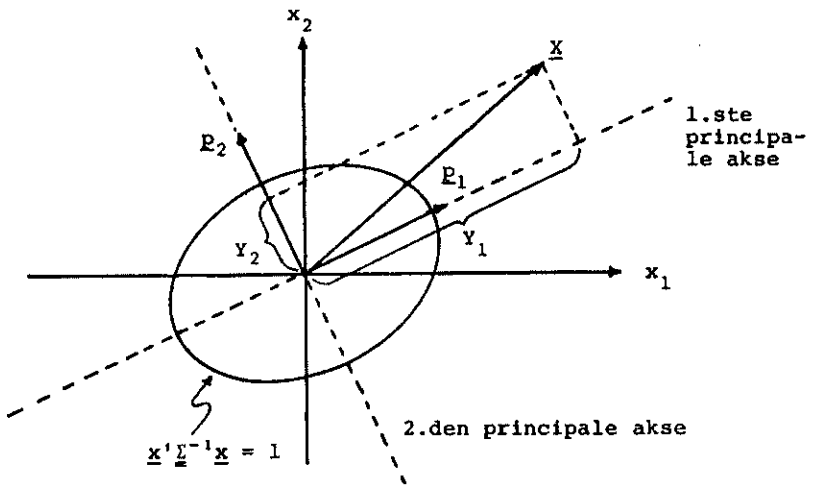
DEFINITION 8.1. Ved den i**'te principale akse** for $\mathbf{X}$ forstås retningen hørende til egenvektoren $\mathbf{p}_i$ svarende til den i'te største egenværdi. ▲

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

Vektoren

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

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

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

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 8.1. Vi har ifølge sætningerne 2.5 og 1.10

$$\mathrm{D}(\mathbf{Y}) = \mathrm{D}(\mathbf{P}'\mathbf{X}) = \mathbf{P}'\boldsymbol{\Sigma}\,\mathbf{P} = \boldsymbol{\Lambda}$$
$$\begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \lambda_k \end{pmatrix},$$

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 8.2. Ifølge definitionen p. 105 haves

$$\mathrm{GV}(\mathbf{X}) = \det \boldsymbol{\Sigma}$$

og

$$\mathrm{GV}(\mathbf{Y}) = \det \boldsymbol{\Lambda} = \lambda_1 \cdots \lambda_k,$$

og af relationen p. 38 følger umiddelbart, at $\mathrm{GV}(\mathbf{X}) = \mathrm{GV}(\mathbf{Y})$. ■

Et lignende resultat er

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

$$\sum_i \mathrm{V}(X_i) = \sum_i \mathrm{V}(Y_i)$$

▲

BEVIS 8.3. Da

$$\sum \mathrm{V}(X_i) = \mathrm{tr}\,\boldsymbol{\Sigma}$$

og

$$\sum \mathrm{V}(Y_i) = \mathrm{tr}\,\mathbf{\Lambda}$$

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

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

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

$$\sup_{\substack{\mathbf{b} \perp \mathbf{p}_1, \ldots, \mathbf{p}_{m-1} \\ \|\mathbf{b}\| = 1}} \mathrm{V}(\mathbf{b}'\mathbf{X}) = \lambda_m,$$

og supremum antages for $\mathbf{b} = \mathbf{p}_m$ ▲

BEVIS 8.4. Da

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

og

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

hvorfor

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

er sætningen blot en omformulering af sætning 1.15 p. 36. ■

BEMÆRKNING 8.1. 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 8.1 og 8.3, 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 $\mathbf{X}$ ud fra vektoren

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

Da

$$\mathbf{Y} = \mathbf{P}'\mathbf{X} \Leftrightarrow \mathbf{X} = \mathbf{P}\,\mathbf{Y},$$

vil det være nærliggende at forsøge med

$$\mathbf{X}^* = \mathbf{P}\,\mathbf{Y}^*$$

Vi finder

$$\begin{aligned} \mathrm{D}(\mathbf{X}^*) &= \mathbf{P}\,\mathrm{D}(\mathbf{Y}^*)\mathbf{P}' \\ &= (\mathbf{p}_1 \cdots \mathbf{p}_k) \begin{pmatrix} \lambda_1 & & \cdots & & 0 \\ & \ddots & & & \\ \vdots & & \lambda_m & & \vdots \\ & & & \ddots & \\ 0 & & \cdots & & 0 \end{pmatrix} \begin{pmatrix} \mathbf{p}_1' \\ \vdots \\ \mathbf{p}_k' \end{pmatrix} \\ &= \lambda_1 \mathbf{p}_1 \mathbf{p}_1' + \cdots + \lambda_m \mathbf{p}_m \mathbf{p}_m'. \end{aligned}$$

Spektralfremstillingen af $\boldsymbol{\Sigma}$ er (p. 31)

$$\boldsymbol{\Sigma} = \lambda_1 \mathbf{p}_1 \mathbf{p}_1 + \cdots + \lambda_m \mathbf{p}_m \mathbf{p}_m' + \cdots + \lambda_k \mathbf{p}_k \mathbf{p}_k',$$

hvorfor

$$\boldsymbol{\Sigma} - \mathrm{D}(\mathbf{X}^*) = \lambda_{m+1} \mathbf{p}_{m+1} \mathbf{p}_{m+1}' + \cdots + \lambda_k \mathbf{p}_k \mathbf{p}_k'.$$

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 Σ 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. [3].

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

$$\mathbf{X}_i = \begin{pmatrix} X_{1\,i} \\ \vdots \\ X_{k\,i} \end{pmatrix}, \qquad i = 1, \ldots, n.$$

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

$$H_0 : \lambda_1 \geq \cdots \geq \lambda_m \geq \lambda_{m+1} = \cdots = \lambda_k$$

mod alternativet, at der forekommer et skarpt ulighedstegn blandt de $k - m$ sidste egenværdier.

Hvis vi arbejder med den estimerede dispersionsmatrix $\hat{\mathbf{\Sigma}}$, bliver teststørrelsen

$$Z_1 = -n' \log_e \frac{\det \hat{\mathbf{\Sigma}}}{\hat{\lambda}_1 \cdots \hat{\lambda}_m \cdot \hat{\lambda}^{k-m}} = -n' \log_e \frac{\hat{\lambda}_{m+1} \cdots \hat{\lambda}_k}{\hat{\lambda}^{k-m}},$$

hvor

$$n' = n - m - \frac{1}{6}(2(k-m) + 1 + \frac{2}{k-m}),$$

og

$$\hat{\lambda} = (\operatorname{tr}\hat{\mathbf{\Sigma}} - \hat{\lambda_1} - \cdots - \hat{\lambda}_m)/(k-m) = (\hat{\lambda}_{m+1} + \cdots + \hat{\lambda}_k)/(k-m).$$

Det kritiske område ved test på niveau α er approximativt

$$\{(\mathbf{x}_1, \cdots, \mathbf{x}_n) | z_1 > \chi^2(\frac{1}{2}(k-m+2)(k-m-1))_{1-\alpha}\}.$$

Regner vi i stedet på den estimerede **korrelationsmatrix** $\hat{\mathbf{R}}$, fås kriteriet

$$Z_2 = -n \log_e \frac{\det \hat{\mathbf{R}}}{\hat{\lambda}_1 \cdots \hat{\lambda}_m \cdot \hat{\lambda}^{k-m}} = -n \log_e \frac{\hat{\lambda}_{m+1} \cdot \hat{\lambda}_k}{\hat{\lambda}^{k-m}},$$

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

$$\{\mathbf{x}_1, \ldots, \mathbf{x}_n | z_2 > \chi^2(\frac{1}{2}(k-m+2)(k-m-1))_{1-\alpha}\}.$$

Det må dog her indskærpes, at denne approximation er langt dårligere end den tilsvarende for dispersionsmatricen.

En diskussion af ovennævnte tests kan findes i [23].

Vi giver nu et eksempel.

EKSEMPEL 8.1. Eksemplet er baseret på et eksempel fra [10] 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 komponentanalyse 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

$$\hat{\Sigma} = \begin{bmatrix} 5.400 & 3.260 & 0.779 & 6.391 & 2.155 & 3.035 & -1.996 \\ 3.260 & 5.846 & 1.465 & 6.083 & 1.312 & 2.877 & -2.370 \\ 0.779 & 1.465 & 2.774 & 2.204 & -3.839 & -5.167 & -1.740 \\ 6.391 & 6.083 & 2.204 & 9.107 & 1.610 & 2.782 & -3.283 \\ 2.155 & 1.312 & -3.839 & 1.610 & 10.710 & 14.770 & 2.252 \\ 3.035 & 2.877 & -5.167 & 2.782 & 14.770 & 20.780 & 2.622 \\ -1.996 & -2.370 & -1.740 & -3.283 & 2.252 & 2.622 & 2.594 \end{bmatrix}$$

Dernæst bestemmes egenvektorerne og egenværdierne for $\hat{\Sigma}$. 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	$\hat{\mathbf{p}}_1$	$\hat{\mathbf{p}}_2$	$\hat{\mathbf{p}}_3$	$\hat{\mathbf{p}}_4$	$\hat{\mathbf{p}}_5$	$\hat{\mathbf{p}}_6$	$\hat{\mathbf{p}}_7$
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

$$Y_1 = 0.164X_1 + \ldots + 0.546X_5 + 0.768X_6 + 0.073X_7$$

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

$$Y_2 = 0.422X_1 + 0.447X_2 + 0.257X_3 + 0.650X_4 + \cdots - 0.313X_7,$$

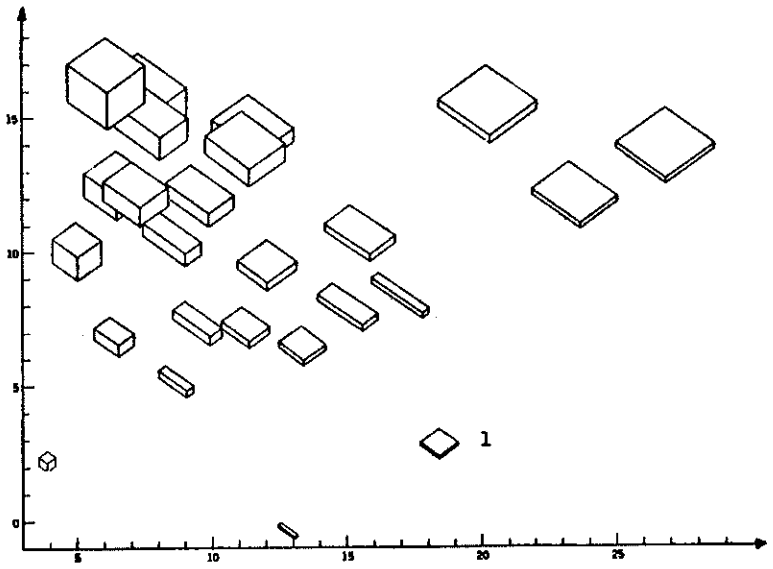
er meget stor, må en eller flere af $X_1, \ldots, X_4$ være store og X_7 lille. Nu gælder det, at terningen er den kasse, der for et givet volumen har den mindste overflade. Hvis en kasse derfor afviger meget fra en terning, vil den have stor X_7-værdi, og dette vil medvirke til en kraftig reduktion 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.

$$\begin{aligned} Y_1 &= 0.164\,3.760 + \cdots + 0.073\,4.782 = 18.18 \\ Y_2 &= 0.422\,3.760 + \cdots - 0.313\,9.782 = 2.15. \end{aligned}$$

I punktet med koordinater $(18.18, 2.15)$ er der dernæst tegnet et billede af kasse nr. 1, etc.



Figur 8.1:

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

$$Y_3 = 0.645X_1 - 0.713X_2 + \cdots + 0.065X_7.$$

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 [2] 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:

$$\hat{\mathbf{R}} = \begin{bmatrix} 1.00 & 0.32 & 0.41 & -0.31 \\ 0.32 & 1.00 & 0.68 & -0.38 \\ 0.41 & 0.68 & 1.00 & -0.36 \\ -0.31 & -0.38 & -0.36 & 1.00 \end{bmatrix}.$$

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

$$\hat{\lambda}_1 = 2.25; \quad \hat{\mathbf{p}}_1 = \begin{bmatrix} 0.43 \\ 0.55 \\ 0.57 \\ -0.44 \end{bmatrix}$$

$$\hat{\lambda}_2 = 0.74; \quad \hat{\mathbf{p}}_2 = \begin{bmatrix} -0.66 \\ 0.49 \\ 0.37 \\ 0.44 \end{bmatrix}$$

$$\hat{\lambda}_3 = 0.70; \quad \hat{\mathbf{p}}_3 = \begin{bmatrix} 0.60 \\ -0.02 \\ 0.16 \\ 0.78 \end{bmatrix}$$

$$\hat{\lambda}_4 = 0.31; \quad \hat{\mathbf{p}}_4 = \begin{bmatrix} -0.14 \\ -0.68 \\ 0.72 \\ -0.06 \end{bmatrix}$$

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

$$Z = -n \log \frac{0.74 \cdot 0.70 \cdot 0.31}{[(0.74 + 0.70 + 0.31)/3]^3} = 0.2120n,$$

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 $0.21n$ 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 $\mathbf{X}$

$$\mathbf{X} \in \mathrm{N}_{p+q}(\mu, \mathbf{\Sigma}),$$

hvor $p \leq q$, og hvor $\mathbf{X}$ og parametrene er spaltet som følger:

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \quad \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix}.$$

Hvis vi på basis af n målinger af $\mathbf{X}$ ønsker at undersøge, om $\mathbf{X}_1$ og $\mathbf{X}_2$ er uafhængige, kan dette som anført i kapitel 6 gøres ved at undersøge

$$\frac{\det(\mathbf{S})}{\det(\mathbf{S}_{11})\det(\mathbf{S}_{22})},$$

der er $U_{p,q,n-l-q}$ fordelt under H_0. Vi vil nu prøve at anskue problemet fra en lidt anden synsvinkel. Vi betragter to endimensionale variable U og V givet ved

$$U = \mathbf{a}'\mathbf{X}_1 \quad \text{og} \quad V = \mathbf{b}'\mathbf{X}_2.$$

Da er

$$D\begin{pmatrix} U \\ V \end{pmatrix} = \begin{pmatrix} \mathbf{a}' \\ \mathbf{b}' \end{pmatrix} \Sigma(\mathbf{a},\mathbf{b}) = \begin{bmatrix} \mathbf{a}'\boldsymbol{\Sigma}_{11}\mathbf{a} & \mathbf{a}'\boldsymbol{\Sigma}_{12}\mathbf{b} \\ \mathbf{b}'\boldsymbol{\Sigma}_{21}\mathbf{a} & \mathbf{b}'\boldsymbol{\Sigma}_{22}\mathbf{b} \end{bmatrix},$$

og korrelationen mellem U og V er

$$\rho(\mathbf{U},\mathbf{V}) = \frac{\mathbf{a}'\boldsymbol{\Sigma}_{12}\mathbf{b}}{\sqrt{\mathbf{a}'\boldsymbol{\Sigma}_{11}\mathbf{a}\,\mathbf{b}'\boldsymbol{\Sigma}_{22}\mathbf{b}}}.$$

Nu er åbenbart

$$\boldsymbol{\Sigma}_{12} = 0 \Leftrightarrow \forall \mathbf{a},\mathbf{b}: \quad \rho(\mathbf{a},\mathbf{b}) = 0.$$

Acceptområdet for hypotesen $\rho(\mathbf{a},\mathbf{b}) = 0$ er af formen (jvf. kapitel 2)

$$r^2(\mathbf{a},\mathbf{b}) \leq r_\beta^2,$$

hvor $r(\mathbf{a},\mathbf{b})$ er den empiriske korrelationskoefficient, og r_β^2 er en passende fraktil i nulhypotesefordelingen. Vi får derfor accepteret $\boldsymbol{\Sigma}_{12} = 0$, såfremt

$$\forall \mathbf{a},\mathbf{b}: \quad r^2(\mathbf{a},\mathbf{b}) \leq r_\beta^2,$$

hvilket helt klart er ensbetydende med, at

$$\max_{\mathbf{a},\mathbf{b}} r^2(\mathbf{a},\mathbf{b}) \leq r_\beta^2.$$

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 $\mathbf{X}_1$'erne og $\mathbf{X}_2$'erne, således at disse kombinationer er uafhængige af de foregående, etc. Helt stringent har vi

DEFINITION 8.3. Lad $\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}$ være en stokastisk variabel, hvor $\mathbf{X}_1$ har p komponenter og $\mathbf{X}_2$ q komponenter $(p \leq q)$. **Det r'te par kanoniske variable** er det par af linearkombinationer $U_r = \alpha_r' \mathbf{X}_1$ og $V_r = \beta_r' \mathbf{X}_2$, 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 $\mathrm{D}(X) = \boldsymbol{\Sigma}$ være spaltet analogt

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

Da er den r'te kanoniske korrelation lig den r'te største rod λ_r af

$$\det \begin{pmatrix} -\lambda \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & -\lambda \boldsymbol{\Sigma}_{22} \end{pmatrix} = 0,$$

og koefficienterne i det r'te par kanoniske variable tilfredsstiller

(i) $$\begin{pmatrix} -\lambda \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & -\lambda \boldsymbol{\Sigma}_{22} \end{pmatrix} \begin{pmatrix} \alpha_r \\ \beta_r \end{pmatrix} = \mathbf{0}$$

(ii) $\alpha_r' \boldsymbol{\Sigma}_{11} \alpha_r = 1$

(iii) $\beta_r' \boldsymbol{\Sigma}_{22} \beta_r = 1.$

▲

BEVIS 8.5. Der er tale om et maksimaliseringsproblem under bibetingelser, og man kan komme igennem ved at anvende en Lagrange-multiplikator-teknik, se f.eks. [3]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

$$\begin{aligned}(\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} - \lambda_r^2\boldsymbol{\Sigma}_{11})\alpha_r &= \mathbf{0} \\ \det(\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} - \lambda_r^2\boldsymbol{\Sigma}_{11}) &= 0\end{aligned}$$

respektive

$$\begin{aligned}(\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} - \lambda_r^2\boldsymbol{\Sigma}_{22})\beta_r &= \mathbf{0} \\ \det(\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} - \lambda_r^2\boldsymbol{\Sigma}_{22}) &= 0\end{aligned}$$

▲

BEVIS 8.6. Forbigås, se f.eks. [3]. ■

Vedrørende **estimationen** er der intet særligt at tilføje. Indsættes maximum likelihood-skøn for $\boldsymbol{\Sigma}$ i de foregående sætninger, fås maximum likelihood-skøn for parametrene. Hyppigst vil man nok indsætte det centrale skøn $\mathbf{S}$, 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

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_2 \end{bmatrix},$$

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

$$\mathbf{X} = \mathbf{A}\,\mathbf{F} + \mathbf{G},$$

d.v.s. skrevet ud

$$\begin{bmatrix} X_1 \\ \vdots \\ X_k \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{k1} & \cdots & a_{km} \end{bmatrix} \cdot \begin{bmatrix} F_1 \\ \vdots \\ F_m \end{bmatrix} + \begin{bmatrix} G_1 \\ \vdots \\ G_k \end{bmatrix}.$$

Her benævnes $\mathbf{F}$ 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.

$\mathbf{A}$-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_1 angiver rumlig intelligens og F_m verbal do., og at X_1 er resultatet af en prøve af rumgeometrisk tilsnit og X_k resultatet af en læseprøve, ja da vil man selvsagt have, at a_{11} er stor og a_{1m} lille og omvendt, at a_{k1} 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 $\mathbf{G}$ kaldes vektoren af **unikke 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 $\mathbf{X}$ som $\mathbf{F}$ og $\mathbf{G}$ antages at være stokastiske. Der er derfor **ikke** tale om en generel lineær model med parametre $F_1, \ldots, F_m$.

For at gøre denne forskel helt tydelig vil vi derfor anføre modellen i det tilfælde, hvor vi har flere observationer $\mathbf{X}_1, \ldots, \mathbf{X}_n$. Vi har da de n modeller

$$\begin{bmatrix} X_{1i} \\ \vdots \\ X_{ki} \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & & \vdots \\ a_{k1} & \cdots & a_{km} \end{bmatrix} \begin{bmatrix} F_{1i} \\ \vdots \\ F_{mi} \end{bmatrix} + \begin{bmatrix} G_{1i} \\ \vdots \\ G_{ki} \end{bmatrix},$$

hvor vi bemærker, at det er $\mathbf{F}_i$ og $\mathbf{G}_i$, der ændres med observationerne $\mathbf{X}_i$.

Vi kan samle ovenstående modeller til

$$\begin{bmatrix} X_{11} \cdots X_{1n} \\ \vdots \quad \vdots \\ X_{k1} \cdots X_{kn} \end{bmatrix} = \begin{bmatrix} a_{11} \cdots a_{1m} \\ \vdots \quad \vdots \\ a_{k1} \cdots a_{km} \end{bmatrix} \begin{bmatrix} F_{11} \cdots F_{1n} \\ \vdots \quad \vdots \\ F_{m1} \cdots F_{mn} \end{bmatrix} + \begin{bmatrix} G_{11} \cdots G_{1n} \\ \vdots \quad \vdots \\ G_{k1} \cdots G_{kn} \end{bmatrix}.$$

Det forudsættes at $\mathbf{F}$ og $\mathbf{G}$ er ukorrelerede, og at

$$\mathrm{D}(\mathbf{F}) = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & 1 \end{pmatrix} = \mathbf{I} = \mathbf{I}_m,$$

og

$$\mathrm{D}(\mathbf{G}) = \begin{pmatrix} \delta_1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \delta_k \end{pmatrix} = \mathbf{\Delta}.$$

Endvidere forudsættes, at observationerne er standardiseret på en sådan måde, at $\mathrm{V}(X_i) = 1, \forall i$, d.v.s. at dispersionsmatricen for $\mathbf{X}$ er lig dens korrelationsmatrix, som benævnes

$$\mathrm{D}(\mathbf{X}) = \mathbf{R} = \begin{pmatrix} 1 & \cdots & r_{1k} \\ \vdots & & \vdots \\ r_{k1} & \cdots & 1 \end{pmatrix}.$$

Af den oprindelige faktorligning fås ved hjælp af sætning 2.5 p. 60, at

$$\mathbf{R} = \mathbf{A}\,\mathbf{A}' + \mathbf{\Delta}.$$

Heraf udledes specielt, at vi for $j = 1, \ldots, k$ har

$$\mathrm{V}(X_j) = a_{j1}^2 + \cdots + a_{jm}^2 + \delta_j = 1.$$

Her indføres betegnelsen

$$h_j^2 = a_{j1}^2 + \cdots + a_{jm}^2, \qquad j = 1, \ldots, k.$$

Disse størrelser benævnes **kommunaliteter** (communalities), og h_j^2 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 [11], at

$$h_j^2 = a_{j1}^2 + \cdots + a_{jm}^2 \geq r_{j|1\ldots k}^2,$$

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 $\mathbf{A}$. Vi finder

$$\mathbf{A}\,\mathbf{A}' = \mathbf{R} - \boldsymbol{\Delta}.$$

Diagonalelementerne i denne matrix er

$$1 - \delta_j = h_j^2, \qquad j = 1, \ldots, k.$$

Disse kender vi ikke, men vi kan eventuelt estimere dem ved de multiple korrelationskoefficienters kvadrater. Indsættes disse, fås en matrix

$$\mathbf{V} = \begin{bmatrix} r_{1|2\cdots k}^2 & \cdots & r_{1k} \\ \vdots & & \vdots \\ r_{k1} & \cdots & r_{k|1\cdots k-1}^2 \end{bmatrix},$$

hvis elementer uden for diagonalen er lig den oprindelige korrelationsmatrix $\mathbf{R}$'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" $\mathbf{V}$. $\mathbf{V}$ kan således f.eks. være den ovenfor nævnte.

Vi benævner $\mathbf{V}$'s egenværdier og tilhørende normerede, ortogonale egenvektorer

$$\lambda_1 \geq \cdots \geq \lambda_k,$$

henholdsvis

$$\mathbf{p}_1, \ldots, \mathbf{p}_k.$$

Sætter vi

$$\mathbf{P} = (\mathbf{p}_1, \ldots, \mathbf{p}_k),$$

har vi ifølge sætning 1.10 p. 30, at

$$\mathbf{P}'\mathbf{V}\,\mathbf{P} = \mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \lambda_k \end{pmatrix}.$$

Da $\mathbf{P}$ er ortogonal, fås

$$\mathbf{V} = \mathbf{P}\,\mathbf{\Lambda}\,\mathbf{P}' = (\mathbf{P}\,\mathbf{\Lambda}^{\frac{1}{2}})(\mathbf{P}\,\mathbf{\Lambda}^{\frac{1}{2}})',$$

hvor

$$\mathbf{\Lambda}^{\frac{1}{2}} = \begin{pmatrix} \sqrt{\lambda_1} & \cdots & 0 \\ \vdots & & \vdots \\ 0 & \cdots & \sqrt{\lambda_k} \end{pmatrix}.$$

Vi definerer nu

$$\mathbf{\Lambda}_*^{\frac{1}{2}} = \begin{pmatrix} \sqrt{\lambda_1} & \cdots & 0 \\ & & \vdots \\ \vdots & & \sqrt{\lambda_m} \\ & & \vdots \\ 0 & \cdots & 0 \end{pmatrix}.$$

d.v.s. $\mathbf{\Lambda}_*^{\frac{1}{2}}$ består af de m første søjler i $\mathbf{\Lambda}^{\frac{1}{2}}$ svarende til de m største egenværdier. Vi ser da

$$\begin{aligned}(\mathbf{P}\,\mathbf{\Lambda}_*^{\frac{1}{2}})(\mathbf{P}\,\mathbf{\Lambda}_*^{\frac{1}{2}})' &= \mathbf{P}\,\mathbf{\Lambda}_*^{\frac{1}{2}}\mathbf{\Lambda}_*^{\frac{1}{2}\prime}\mathbf{P}' \\ &= \mathbf{P}\begin{pmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \lambda_m & \vdots \\ 0 & \cdots & 0 \end{pmatrix}\mathbf{P}' \\ &\simeq \mathbf{V},\end{aligned}$$

jvf. de analoge betragtninger p. 348.

Da $\mathbf{V}$ var et skøn over $\mathbf{A}\,\mathbf{A}'$, har vi derfor

$$\mathbf{A}\,\mathbf{A}' \simeq (\mathbf{P}\,\mathbf{\Lambda}_*^{\frac{1}{2}})(\mathbf{P}\,\mathbf{\Lambda}_*^{\frac{1}{2}})',$$

hvorfor det vil være naturligt at vælge $\mathbf{P}\,\mathbf{\Lambda}_*^{\frac{1}{2}}$ som skøn over $\mathbf{A}$.Denne løsning kaldes "principal faktor"-løsningen til vort estimationsproblem.

Vi samler overvejelserne i følgende

SÆTNING 8.7. Vi betragter faktormodellen $\mathbf{X} = \mathbf{A}\,\mathbf{F} + \mathbf{G}$, hvor $\mathbf{X}$ er k-dimensional og $\mathbf{F}$ m-dimensional. $\mathbf{X}$'s korrelationsmatrix betegnes $\mathbf{R}$, og $\mathbf{V}$ er den matrix, der fremkommer ved at erstatte 1-tallene i $\mathbf{R}$'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 faktor løsningen** til estimationsproblemet er da

$$\mathbf{P}\,\mathbf{\Lambda}_*^{\frac{1}{2}} = (\sqrt{\lambda_1}\mathbf{p}_1, \ldots, \sqrt{\lambda_m}\mathbf{p}_m),$$

hvor λ_i, $i = 1, \ldots, m$, er de m største egenværdier til $\mathbf{V}$, og hvor $\mathbf{p}_i$, $i = 1, \ldots, m$, er de tilsvarende normerede egenvektorer. ▲

BEMÆRKNING 8.2. 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

$$\mathbf{A}\,\mathbf{A}' \simeq (\mathbf{P}\,\mathbf{\Lambda}_*^{\frac{1}{2}})(\mathbf{P}\,\mathbf{\Lambda}_*^{\frac{1}{2}})'$$

Hvis nu $\mathbf{Q}$ er en vilkårlig $m \times m$ ortogonal matrix, d.v.s.: $\mathbf{Q}\,\mathbf{Q}' = \mathbf{I}$, har vi

$$\begin{aligned}(\mathbf{P}\,\mathbf{\Lambda}_*^{\frac{1}{2}}\mathbf{Q})(\mathbf{P}\,\mathbf{\Lambda}_*^{\frac{1}{2}}\mathbf{Q})' &= (\mathbf{P}\,\mathbf{\Lambda}_*^{\frac{1}{2}})\mathbf{Q}\,\mathbf{Q}'(\mathbf{P}\,\mathbf{\Lambda}_*^{\frac{1}{2}})' \\ &= (\mathbf{P}\,\mathbf{\Lambda}_*^{\frac{1}{2}})(\mathbf{P}\,\mathbf{\Lambda}_*^{\frac{1}{2}})' \\ &= \mathbf{A}\,\mathbf{A}'.\end{aligned}$$

Dette indebærer altså, at vi kan få vilkårligt mange estimater for $\mathbf{A}$-matricen ved at multiplicere principal faktor løsningen med en ortogonal matrix.

Problemet er så blot, hvorledes man hensigtsmæssigt vælger $\mathbf{Q}$-matricen. Hovedprincippet er, at man ønsker, at $\mathbf{A}$-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 $\mathbf{Q}$ således, at størrelsen

$$\sum_j m \left\{ \sum_i \left(\frac{a_{ij}^2}{h_i^2} \right)^2 - \frac{1}{m} \left[\sum_i \left(\frac{a_{ij}^2}{h_i^2} \right) \right]^2 \right\}$$

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. 351

finder vi

$$\hat{\mathbf{R}} = \begin{bmatrix} 1.000 & 0.580 & 0.201 & 0.911 & 0.283 & 0.287 & -0.533 \\ 0.580 & 1.000 & 0.364 & 0.834 & 0.166 & 0.261 & -0.609 \\ 0.201 & 0.364 & 1.000 & 0.439 & -0.704 & -0.681 & -0.649 \\ 0.911 & 0.834 & 0.439 & 1.000 & 0.163 & 0.202 & -0.676 \\ 0.283 & 0.166 & -0.704 & 0.163 & 1.000 & 0.990 & 0.427 \\ 0.287 & 0.261 & -0.681 & 0.202 & 0.990 & 1.000 & 0.357 \\ -0.533 & -0.609 & -0.649 & -0.676 & 0.427 & 0.357 & 1.000 \end{bmatrix}$$

Fuldstændigt i analogi med fremgangsmåden i eksempel 8.1 bestemmes dernæst egenværdier og -vektorer for $\hat{\mathbf{R}}$. Vi finder

Egenværdi $\hat{\lambda}_i, 1, \ldots, 7$	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 $(\sqrt{\hat{\lambda}_1}\hat{\mathbf{p}}_1, \sqrt{\hat{\lambda}_2}\hat{\mathbf{p}}_2)$, hvor

$$\begin{pmatrix} \sqrt{\hat{\lambda}_1}\hat{\mathbf{p}}_1' \\ \sqrt{\hat{\lambda}_2}\hat{\mathbf{p}}_2' \end{pmatrix} = \begin{pmatrix} 0.747 & 0.795 & 0.710 & 0.910 & -0.235 & -0.178 & -0.886 \\ 0.491 & 0.373 & 0.596 & -0.389 & -0.963 & -0.971 & 0.218 \end{pmatrix}.$$

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

$$\hat{\mathbf{h}}^{2\prime} = [\ 0.798 \quad 0.771 \quad 0.860 \quad 0.979 \quad 0.983 \quad 0.976 \quad 0.833\ ],$$

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. 359). Vi finder

$$\delta' = [\ 0.202 \quad 0.229 \quad 0.140 \quad 0.021 \quad 0.017 \quad 0.024 \quad 0.167\ ].$$

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

$$\hat{\mathbf{Z}} = \hat{\mathbf{R}} - \hat{\mathbf{A}}\hat{\mathbf{A}}',$$

og får som mål for faktorernes evne til at beskrive den oprindelige variabilitet i materialet

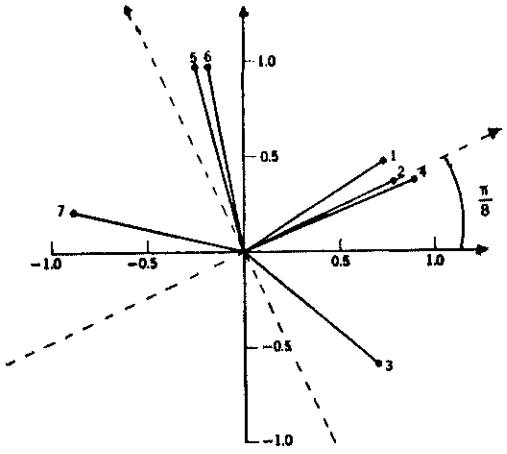
$$\hat{\mathbf{Z}} = \begin{bmatrix} 0.202 & -0.196 & -0.037 & 0.041 & -0.914 & -0.057 & 0.021 \\ -0.196 & 0.229 & 0.071 & -0.035 & -0.006 & 0.041 & 0.015 \\ -0.037 & 0.021 & 0.140 & 0.024 & 0.037 & 0.025 & 0.111 \\ 0.041 & -0.035 & 0.024 & 0.021 & 0.002 & -0.013 & 0.046 \\ -0.014 & -0.006 & 0.037 & 0.002 & 0.017 & 0.012 & 0.009 \\ -0.057 & 0.041 & 0.025 & -0.013 & 0.012 & 0.024 & -0.013 \\ 0.021 & 0.015 & 0.111 & 0.046 & 0.009 & -0.013 & 0.167 \end{bmatrix}.$$

Jo mere $\hat{\mathbf{Z}}$ afviger fra $\mathbf{Q}$-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. 365) $\hat{a}_{ij}$ 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. [12] eller [6]).

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 [12] og [6].

8.3.4 Beregning af faktorværdier (factor scores)

Hvis vi i ovenstående eksempel 8.3 ønsker at lave et diagram analogt til det p. 353, 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 $\mathbf{G}$.

Vi har modellen (jvf. p. 359)

$$\mathbf{X} = \mathbf{A}\,\mathbf{F} + \mathbf{G},$$

hvor

$$\begin{aligned} \mathrm{D}(\mathbf{F}) &= \mathbf{I} \\ \mathrm{D}(\mathbf{G}) &= \boldsymbol{\Delta}, \end{aligned}$$

og hvor $\mathbf{F}$ og $\mathbf{G}$ er ukorrelerede.

Derfor er

$$\mathrm{D}\begin{pmatrix} \mathbf{X} \\ \mathbf{F} \end{pmatrix} = \begin{pmatrix} \mathbf{A}\,\mathbf{A}' + \boldsymbol{\Delta} & \mathbf{A} \\ \mathbf{A}' & \mathbf{I} \end{pmatrix}.$$

Da som tidligere nævnt

$$\mathrm{Cov}(X_i, F_j) = a_{ij},$$

følger, at matricerne uden for diagonalen netop er $\mathbf{A}$-matricen, respektive dens transponerede.

Skønnet over denne dispersionsmatrix er

$$\begin{bmatrix} \hat{\mathbf{A}}\,\hat{\mathbf{A}}' + \hat{\boldsymbol{\Delta}} & \hat{\mathbf{A}} \\ \hat{\mathbf{A}}' & \mathbf{I} \end{bmatrix}.$$

Den betingede fordeling af $\mathbf{F}$ for givet $\mathbf{X}$ har - såfremt de underliggende fordelinger er normale - som middelværdi

$$\mu_F + \mathbf{A}'(\mathbf{A}\,\mathbf{A}' + \boldsymbol{\Delta})^{-1}(\mathbf{x} - \mu_x)$$

(jvf. afsnit 2.2.3).

Da vi foretager vore beregninger på de standardiserede x-værdier, er det rimeligt at antage, at $\mu_x = \mathbf{0}$. Niveauet for faktorskalaerne er arbitrært, men det er sædvane også at sætte det lig 0, således at vi får udtrykket

$$\mathbf{A}'(\mathbf{A}\,\mathbf{A}' + \boldsymbol{\Delta})^{-1}\mathbf{x}$$

for den betingede middelværdi af $\mathbf{F}$.

Som estimat af den i'te måling af $\mathbf{X}_i$'s faktorværdi har vi da

$$\hat{\mathbf{F}}_i = \hat{\mathbf{A}}'(\hat{\mathbf{A}}\,\hat{\mathbf{A}}' + \hat{\boldsymbol{\Delta}})^{-1}\mathbf{X}_i. \tag{8.1}$$

Nu vil $\mathbf{A}$-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

$$(\mathbf{A}\mathbf{A}' + \boldsymbol{\Delta})^{-1}\mathbf{A} = \boldsymbol{\Delta}^{-1}\mathbf{A}(\mathbf{I} + \mathbf{A}'\boldsymbol{\Delta}^{-1}\mathbf{A})^{-1},$$

som giver

$$\hat{\mathbf{F}}_i = (\mathbf{I} + \hat{\mathbf{A}}'\hat{\boldsymbol{\Delta}}^{-1}\hat{\mathbf{A}})^{-1}\hat{\mathbf{A}}'\hat{\boldsymbol{\Delta}}^{-1}\mathbf{X}_i. \tag{8.2}$$

Identitetens gyldighed fremgår af følgende relationer

$$\begin{aligned} & (\mathbf{A}\mathbf{A}' + \boldsymbol{\Delta})^{-1}\mathbf{A} &&= \boldsymbol{\Delta}^{-1}\mathbf{A}(\mathbf{I} + \mathbf{A}'\boldsymbol{\Delta}^{-1}\mathbf{A})^{-1} \\ \Leftrightarrow \quad & \mathbf{A} &&= (\mathbf{A}\mathbf{A}' + \boldsymbol{\Delta})\boldsymbol{\Delta}^{-1}\mathbf{A}(\mathbf{I} + \mathbf{A}'\boldsymbol{\Delta}^{-1}\mathbf{A})^{-1} \\ & &&= \mathbf{A}(\mathbf{A}'\boldsymbol{\Delta}^{-1}\mathbf{A} + \mathbf{I})(\mathbf{I} + \mathbf{A}'\boldsymbol{\Delta}^{-1}\mathbf{A})^{-1}, \end{aligned}$$

og den sidste relation er jo trivielt opfyldt.

Nu er $\mathbf{I} + \mathbf{A}'\boldsymbol{\Lambda}^{-1}\mathbf{A}$ en $m \times m$ 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 $(\mathbf{A}\,\mathbf{A}' + \mathbf{\Lambda})$ en $k \times k$ 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

$$\mathbf{R} = \mathbf{A}\mathbf{A}' + \mathbf{\Delta}$$

(jvf. p. 360). Dette giver det med 8.1 ækvivalente udtryk

$$\hat{\mathbf{F}}_i = \hat{\mathbf{A}}'\hat{\mathbf{R}}^{-1}\mathbf{X}_i \tag{8.3}$$

Det må slutteligen præciseres, at der findes en række andre metoder til bestemmelse af faktorværdier, se f.eks. [12] eller [26]. 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. 364, fandt vi en roteret faktorløsning med 2 faktorer. De roterede faktorvægte var

$$\hat{\mathbf{A}} = \begin{bmatrix} 0.878 & 0.168 \\ 0.877 & 0.040 \\ 0.428 & -0.828 \\ 0.990 & 0.011 \\ 0.151 & 0.980 \\ 0.207 & 0.965 \\ -0.735 & 0.540 \end{bmatrix}.$$

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. 360)

$$\hat{h}_j^2 = \hat{a}_{ji}^2 + \hat{a}_{j2}^2 = 1 - \hat{\delta}_j.$$

Vi bemærker, at de her angivne kommunaliteter er lig dem, vi fandt p. 366 for de uroterede faktorer. Dette er alment gyldigt og kan anvendes som et check ved beregningen af de roterede faktorer.

Idet vi har

$$\hat{\mathbf{\Delta}} = \text{diag}(\hat{\delta}_j),$$

d.v.s.

$$\hat{\mathbf{\Delta}}^{-1} = \text{diag}(\frac{1}{\hat{\delta}_j}),$$

bliver

$$(\mathbf{I} + \hat{\mathbf{A}}'\hat{\mathbf{\Delta}}^{-1}\hat{\mathbf{A}})^{-1}\hat{\mathbf{A}}'\hat{\mathbf{\Delta}}^{-1} =$$
$$\begin{bmatrix} 0.0669 & 0.0597 & 0.0593 & 0.7839 & 0.0244 & 0.0510 & -0.0750 \\ -0.0002 & -0.0059 & 0.0655 & -0.0943 & 0.5770 & 0.3641 & 0.0415 \end{bmatrix}$$

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
$\bar{X}_j$	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

$$\mathbf{z} = (-1.4373 \quad -0.4603 \quad -1.0860 \quad -1.2787 \quad 1.3167 \quad 1.4422 \quad 1.5124)',$$

hvor f.eks. den anden værdi fremkommer som

$$z_2 = \frac{3.660 - 4.773}{2.4178} = -0.4603.$$

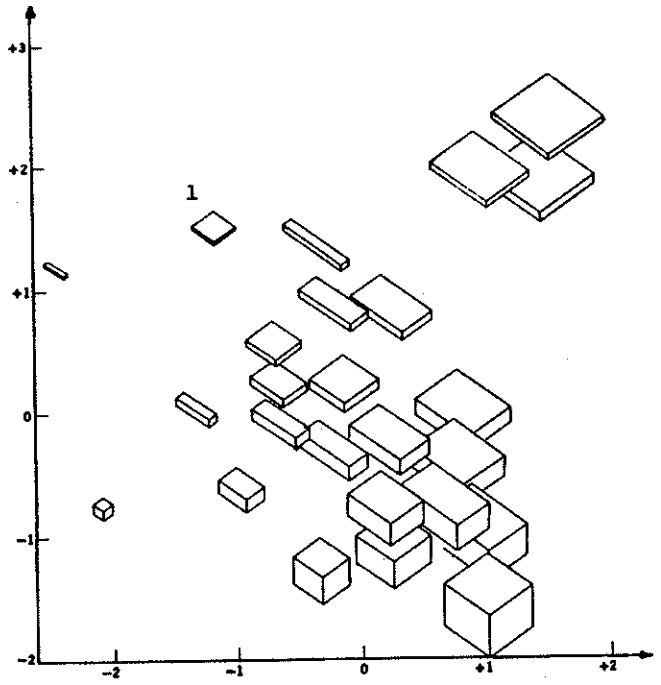
Vi finder nu let faktorværdierne svarende til den første æske som

$$\hat{\mathbf{F}}^1 = (\mathbf{I} + \hat{\mathbf{A}}'\hat{\mathbf{\Delta}}^{-1}\hat{\mathbf{A}})^{-1}\hat{\mathbf{A}}'\hat{\mathbf{\Delta}}^{-1}\mathbf{z} = \begin{pmatrix} -1.20 \\ 1.40 \end{pmatrix}.$$

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. 353).

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 (⅔ 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	0,97	– 0,09	– 0,04	0,95
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	–0,76	– 0,26	0,39	0,79
14. Pct. af de beskæftigede, der er selvstændige	–0,30	0,81	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	0,76	0,20	0,16	0,64
19. " " " " eenfamiliehuse	0,83	0,42	0,22	0,92
20. " " " " tofamiliehuse	0,03	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	0,06	0,96
24. " " værelser pr. person	–0,43	0,77	0,36	0,91
25. " " husstande pr. lejlighed	0,35	0,69	– 0,12	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

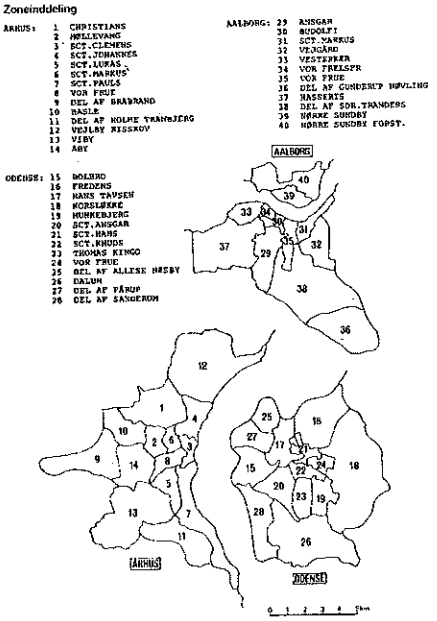


Fig. 1. Den anvendte zoneinddeling af Århus, Odense og Ålborg.
Fig. 1. The 40 zones applied in the analysis, distributed on the three towns of Århus, Odense, and Ålborg.

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,

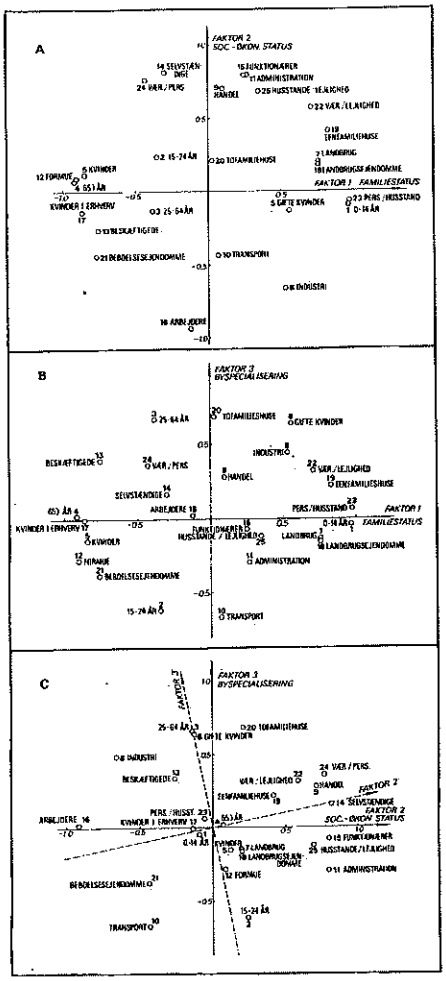


Fig. 2. Diagrammer af sammenhængen mellem faktorvægtene for de tre principale faktorer. Hvert punkt i diagrammerne svarer til en variabel.

Fig. 2. Diagrams showing the interplay between the factor weights for the three principal factors. Each point corresponds to one variable.

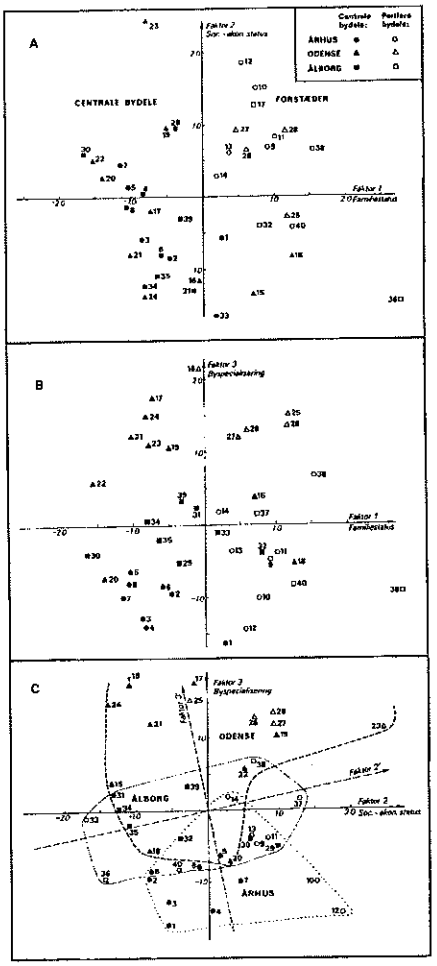


Fig. 3. Diagrammer af sammenhængen mellem faktorværdierne for de tre faktorer. De tre diagrammer viser de tre plane projektioner af det tredimensionale faktorrum. Hvert punkt i diagrammerne svarer til en zone.

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.

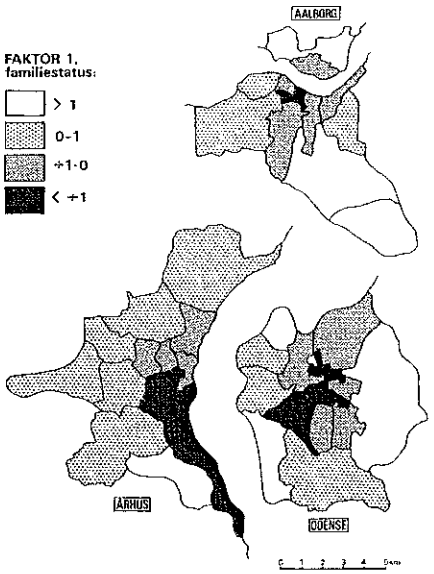


Fig. 4. Kort over faktorværdierne for faktor 1, familiestatus.
Fig. 4. Map showing the factor values for factor 1, family status.

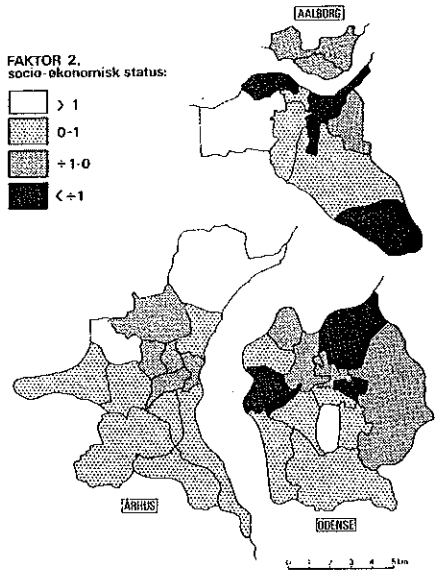


Fig. 5. Kort over faktorværdierne for faktor 2, socio-økonomisk status.
Fig. 5. Map showing the factor values for factor 2, socio-economic status.

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 enfamiliehuse, for husstandsstørrelse og for boligstørrelse, og høje negative faktorvægte for aldersgruppen over 64 år, for folk der er ernæret af for-

mue 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 Alborg 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.

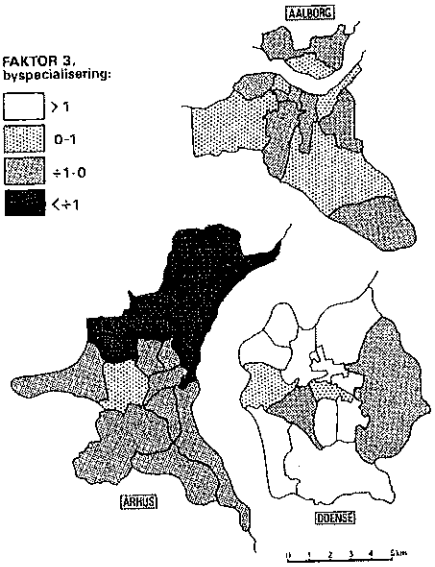


Fig. 6. Kort over faktorværdierne for faktor 3, byspecialisering.
Fig. 6. Map showing the factor values for factor 3, urban specialization.

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. Alborg, 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 $\mathbf{S}$ og har under forudsætning af normalitet af observationerne, at $\mathbf{S}$ er Wishart-fordelt med parametre $(n-1, \frac{1}{n-1}\Sigma)$, hvor Σ er lig $\mathrm{D}(\mathbf{X}_i)$, d.v.s. tætheden er

$$c_1(\det \mathbf{S})^{\frac{1}{2}(n-k-2)}(\det \Sigma)^{-\frac{1}{2}(n-1)} \exp(-\frac{1}{2}(n-1)\operatorname{tr}(\mathbf{S}\,\Sigma^{-1})),$$

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 Σ,

$$\log \mathrm{L}(\Sigma) = -\frac{1}{2}(n-1)\log(\det \Sigma) - \frac{1}{2}(n-1)\operatorname{tr}(\mathbf{S}\,\Sigma^{-1}).$$

Heri introduceres nu den sædvanlige m-faktor model

$$\mathrm{D}(\mathbf{X}) = \Sigma = \mathbf{A}\mathbf{A}' + \Delta,$$

hvor $\mathbf{A}$ og Δ er som i afsnit 8.3.4. Bemærk i øvrigt, at vi her ikke forudsætter, at Σ har et-taller i diagonalen. Dette giver

$$\begin{aligned}\log \mathrm{L}(\mathbf{A}, \Delta) \;=\; & -\frac{1}{2}(n-1)\log(\det(\mathbf{A}\mathbf{A}' + \Delta)) \\ & -\frac{1}{2}(n-1)\operatorname{tr}(\mathbf{S}(\mathbf{A}\mathbf{A}' + \Delta)^{-1}).\end{aligned}$$

Maksimalisering af denne funktion med hensyn til $\mathbf{A}$ og Δ giver ML-løsningen til vores faktoranalyse. Med hensyn til de tekniske problemer, der er involveret heri henvises til [18].

Ved partiel differentiation af logaritmen til likelihood-funktionen kommer man efter lange udregninger og algebraiske manipulationer frem til ligningen

$$\hat{\mathbf{A}} = (\hat{\Delta} + \hat{\mathbf{A}}\hat{\mathbf{A}}')\mathbf{S}^{-1}\hat{\mathbf{A}}, \tag{8.4}$$

se f.eks. [26].

Foretager vi en skalatransformation af $\mathbf{X}$'erne, d.v.s. indfører

$$\mathbf{Z}_i = \mathbf{C}\,\mathbf{X}_i,$$

bliver

$$\mathbf{S}_z = \mathbf{C}\,\mathbf{S}_x\mathbf{C}'$$

hvor z og x som fodtegn angiver, om de forskellige størrelser er beregnet på basis af $\mathbf{Z}_i$ eller $\mathbf{X}_i$'erne. Med samme notationskonvention fås derfor

$$\hat{\mathbf{A}}_z = (\hat{\boldsymbol{\Delta}}_z + \hat{\mathbf{A}}_z\hat{\mathbf{A}}_z')\mathbf{C}'^{-1}\mathbf{S}_x^{-1}\mathbf{C}^{-1}\hat{\mathbf{A}}_z.$$

Ved præ-multiplikation med $\mathbf{C}^{-1}$ fås

$$\mathbf{C}^{-1}\hat{\mathbf{A}}_z = [\mathbf{C}^{-1}\hat{\boldsymbol{\Delta}}_z\mathbf{C}'^{-1} + \mathbf{C}^{-1}\hat{\mathbf{A}}_z(\mathbf{C}^{-1}\hat{\mathbf{A}}_z)']\mathbf{S}_x^{-1}\mathbf{C}^{-1}\hat{\mathbf{A}}_z. \tag{8.5}$$

Ved sammenligning af 8.4 og 8.5 ses, at såfremt $\mathbf{A}$ er en løsning til 8.4, da vil

$$\mathbf{A}_z = \mathbf{C}^{-1}\mathbf{A}$$

være en løsning til 8.5. Med andre ord **medfører en skalering af X'erne (observationerne) med matricen C, at faktorvægtene skaleres med $\mathbf{C}^{-1}$.**

Hvis vi opretholder normalitetsforudsætningen, kan vi **teste, om faktormodellen holder**, d.v.s. teste

$$H_0 : \boldsymbol{\Sigma} = \boldsymbol{\Delta} + \mathbf{A}\mathbf{A}' \quad \text{mod} \quad H_1 : \boldsymbol{\Sigma} \text{ vilkårlig.}$$

Kvotienttestet vil da være ækvivalent med testet givet ved teststørrelsen

$$Z = (n - 1 - \frac{1}{6}(2k + 5) - \frac{2}{3}m)\log_e \frac{|\hat{\boldsymbol{\Delta}} + \hat{\mathbf{A}}\hat{\mathbf{A}}'|}{|\mathbf{S}|}$$

og forkaste for

$$Z > \chi^2(\frac{1}{2}\{(k - m)^2 - k - m\}).$$

Slutteligen skal opmærksomheden henledes på, at der i visse standardprogrammer - f.eks. i HMDP-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 [19]).

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 $\mathbf{X}_1, \ldots, \mathbf{X}_n$, og sætter vi

$$\mathbf{X}' = \begin{bmatrix} X_{11} & \cdots & X_{1n} \\ \vdots & & \vdots \\ X_{k1} & \cdots & X_{kn} \end{bmatrix},$$

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

$$\mathbf{R} = \mathbf{X}'\mathbf{X},$$

jvf. sætning 2.19. **Dualt** kan man selvfølgelig definere

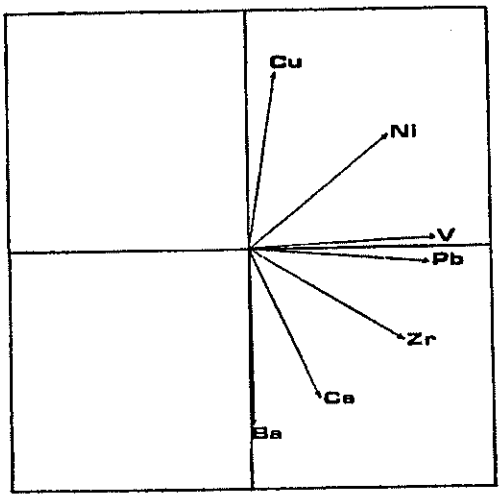
$$\mathbf{Q} = \mathbf{X}\mathbf{X}',$$

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 [22].

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.

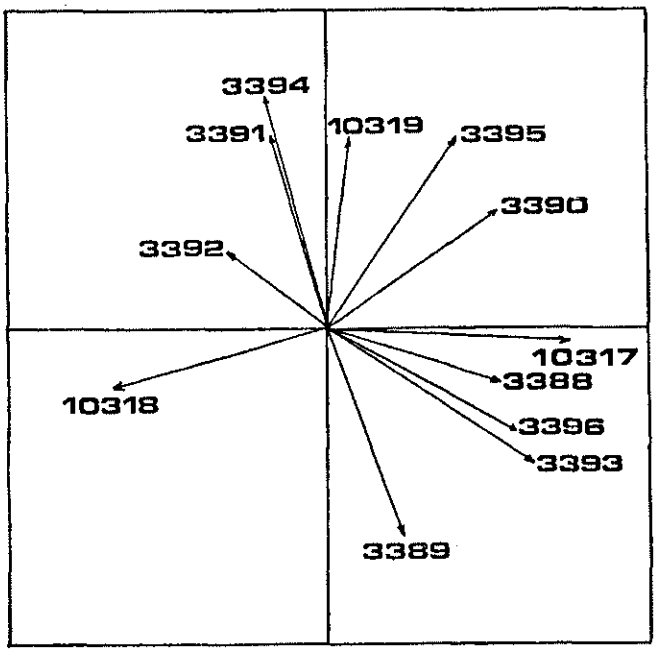
I nedenstående figur er vist de roterede faktorvægte.



Figur 8.2: 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.

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.



Figur 8.3: Faktorvægte i Q-modus analyse.

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 $\mathbf{X}\,\mathbf{X}'$, 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 sddan 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 BMDOlM og BMDO2M fra BMD-systemet.

BMDOlM, 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.

BMDO2M, REGRESSION ON PRINCIPAL COMPONENTS, beregner de samme størrelser som BMDOlM, 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. 364) 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 BMDOBM, 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. 387-387 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. 387.

Litteratur

[1] *Harwell Subroutine Library. Atomic Energy Research Establishment, Harwell 1973.*

[2] AGTERBERG, F. P. *Geomathematics. Mathematical background and geo-science applications.* Elsevier, Amsterdam 1973.

[3] ANDERSON, T. W. *An Introduction to Multivariable Statistical Analysis.* John Wiley & sons, New York 1958.

[4] BENNETT, C. A., AND FRANKLIN, N. L. *Statistical Analysis in Chemistry and the Chemical Industry.* John Wiley & Sons, New York 1954.

[5] BOURBAKI, N. *Algèbre.* Hermann, Paris 1967, ch. 2 Algèbre Lineaire.

[6] CATTELL, R. Factor analysis: An introduction to essentials. i.the purpose and underlying models. ii.the role of factor analysis in research. *Biometrics 21* (1965), 190–215, 405–435.

[7] CORNFIELD, J., AND TUKEY, J. W. Average values of mean squares in factorials. In *The Annals of Mathematical Statistics*, vol. 27. 1956, pp. 907–949.

[8] COX, D. R., AND HINKLEY, D. V. *Theoretical Statistics.* Chapman and Hall, London 1974.

[9] DAVIES, O. L., Ed. *Design and Analysis of Industrial Experiments*, second ed. Oliver and Boyd, London 1967.

[10] DAVIS, J. C. *Statistics and data analysis in geology.* John Wiley, New York 1973.

[11] DWYER, P. S. *The contribution of an orthogonal multiple facto solution to multiple correlation*, vol. 4. John Wiley & Sons, 1939.

[12] HARMAN, H. H. *Modern Factor Analysis*, second ed. The University of Chicago Press, Chicago 1967.

[13] HICKS, C. R. *Fundamental Concepts in the Design of Experiments*, second ed. Holt, Rinehart and Winston, New York 1973.

[14] HOERL, A. E., AND KENNARD, R. W. Ridge regression. biased estimation for nonorthogonal problems. In *Technometrics*, vol. 12,No.1. 1970, pp. 55–67.

[15] JOHNSON, J. R. *An Application of the Design of Experiments in the Surveillance of Ammunition. Proceedings of the First Conference on the Design of Experiments in Army Research, Development and Testing*. 1957.

[16] JOHNSON, N. L., AND LEONE, F. C. *Statistics and Experimental Design in Engineering and the Physical Sciences*, vol. I+II. John Wiley & Sons, New York 1964.

[17] JOHNSON, R. M. On a theorem stated by eckart and young. In *Psychometrika*, vol. 28. 1963, pp. 259–263.

[18] JÖRESKOG, K. G. Some contributions to maximum likelihood factor analysis. In *Psycometrika*, vol. 32. 1967.

[19] KAISER, H. F. *The varimax criterion for analytic rotation in factor analysis*. 1958.

[20] KENDALL, M. G., AND STUART, A. *The Advanced Theory of Statistics*, vol. 2. Charles Griffin & Co., London 1967.

[21] KNUDSEN, J. G. En statistisk analyse af cementstyrke. Eksamensprojekt, DTU,IMSOR, 1975.

[22] LARSEN, P. M. Geokemisk oversigtsprospektering. multivarable statistiske metoders anvendelighed ved interpretation af regionale geokemiske data. Eksamensprojekt, DTU,IMSOR, 1976.

[23] LAWLEY, D. N. *The estimation of factor loadings by the method of maximum likelihood*. 1940.

[24] MARQUARDT, D. W. Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation. In *Technometrics*, vol. 12,No.3. 1970, pp. 591–612.

[25] MARQUARDT, D. W., AND R.D.SNEE. Ridge regression in practice. *The American Statistician 29* (1975), 3–20.

[26] MORRISON, D. F. *Multivariate Statistical Methods*. McGraw-Hill, New York 1967.

[27] PEDERSEN, S. T., AND SKJØTH, P. Statistisk analyse af data fra cementfabrikation. Eksamensprojekt, DTU,IMSOR, 1976.

[28] RAO, C. R. *Estimation and tests of significance in factor analysis*. 1955.

[29] RAO, C. R., AND MITRA, S. K. *Generalized Inverse of Matrices and Its Applications.* John Wiley, New York 1971.

[30] SALOMONSEN, E. Fjernelse af klorider fra forhistorisk jern. Master's thesis, Konservatorskolen, København 1977.

[31] SCHEFFÉ, H. *The Analysis of Variance.* John Wiley & Sons, New York 1959.

[32] SPLIID, H. *User Guide for Stepwise Regression Program REGRGO.* IMSOR, Lyngby 1974.

[33] WILK, M. B., AND KEMPTHORNE, O. Fixed, mixed and random models in the analysis of variance. *Journal of the American Statistical Association 50* (1955), 1144–1167.

[34] WILKINSON, J. H. Error analysis of direct methods of matrix inversion. *Journal of the Association of Computing Machinery 8* (1961), 281–330.

[35] WITTING, H., AND NÖLLE, G. *Angewandte Mathematische Statistik.* B.G.Teubner, Stuttgart 1970.

Indeks

a posteriori fordeling, 310
a priori fordeling, 310
affin afbildning, 7
affin støtte, 66
algebra, *1*
Andersons U, 289
ANOVA, 264
associativ lov, 2

backwards elimination, 184 f, 191
balanceret variansanalyse, 224
beregningsformler, 154
beslutningsfunktion, 310
betinget fordeling, *74*, *93*
bibetingelser, 130

"canonical variables, the first two", 335
central grænseværdisætning, 75
Cholesky faktorisering, 32
cofactor, 15
Cook's D, 168
COVRATIO, 169
Cramér Rao's ulighed, 107
Cramér's sætning, 16

definit, 37
determinant, *14*, 17
DFBETAS, 170
DFFITS, 169
diag, 9
diagonalelement, 9
differentiation af
    kvadratisk form, 47
    linearform, 47
dim, 3
dimension, 3
direkte sum, *5*, 53
discriminant score, 328
diskriminantanalyse, 309 f
    Bayes løsning, 309 f, 327 f, 329 f
    flere normale populationer, 329 f, 334 f
    flere populationer, 327 f
    minimax løsning, 309 f
    to normale populationer, 312 f
    to populationer, 309 f
    ukendte parametre, med, 321 f, 333, 334 f
diskriminantfunktion
    diskriminator, 313, 334
diskriminantværdi, 328
distributive lov, 2
drejning, 32

Eckart-Young's sætning, 34
effekt, 215
egenvektor, *29*, 36
egenvektor m.h.t. matrix, 43
egenværdi, *29*, 36, 46, 47
egenværdi m.h.t. matrix, 43
egenværdiproblem, generelle, *43*
ellipsoide, 40
elliptisk cylinder, 41
empirisk dispersionsmatrix, 76
empirisk generaliseret varians, 105
empirisk partiel korrelation, 88

enhedsmatrix, 31
equimax, 386
estimation af/i
  dispersionsmatrix, 321
  faktor værdi, 368
  dispersionsmatrix, 75, 274, 279, 286 f
  egenværdi i dispersionsmatrix, 349
  faktor vægte, 361 f
euklidisk afstand, 51

factor loading, 359
factor scores, 359
faktor, 215
faktor analyse, 358 f
  principal faktorløsning, 363
  Q-modus analyse, 383 f
  test for model, 382 f
  estimation af vægte, 361 f
  maximum likelihood analyse, 381 f
  rotation, 364 f
faktor vægt, 359
faktor værdi, 359
faktorer, fælles, 359
faktorer, unikke, 359
flerdimensional generel lineær model, 283 f, 285 f
  flerdimensional variansanalyse, 294, 296, 298
  flerdimensionale parametre, *106*
  multipel korrelationskoefficient, 91
  normal fordeling, 75
  partiel korrelationskoefficient, 85
flerdimensional variansanalyse, 294, 298
forventningsværdi, *58*
forward selection, 186 f
funktionel relation, 199

Gauss-Markov's sætning, 115, 285
general lineær model
  flerdimensional, 283
generaliseret varians, 102, 105
generel lineær model, *111*, *141*
  flerdimensional, 283 f
geodæsi, 135

Hotellings T^2
  Tostikprøvesituationen, 322
  enstikprøvesituation, 273 f
  tostikprøvesituation, 279 f, 306 f

idempotent afbildning, 5
idempotent matrix, 11, 47
indre produkt, *51*
Influence Statistics, 165
informationsmatrix, 106
invers matrix, 10, 16, 32, 51
inverst element, 2
isomorfi, 8

kanonisk korrelationskoefficient, 356 f
kanonisk variabel, 355 f
kanoniske korrelationer, 358
kanoniske variable, 358
knude, 213
kommunalitet, 360
kommutativ lov, 2
komplement, 15
konfidensinterval for
  korrelationskoefficient, 89
  partiel korrelationskoefficient, 89
Konfidensintervaller for
  forudsagt værdi, 135
konfidensområdet for
  middelværdi, 277
konfidensområde for
  middelværdi, 280
konjugerede retninger, 53
konjugerede vektorer, 44
konturellipsoide, *69*, 73
koordinater, 4
koordinattransformation, *11*
koordinattransformationsmatrix, 12
korrelationskoefficient, 77, 79
korrelationsmatrix, 61
korrespondanceanalyse, 385
kovarians, *61*
kovariansmatrix, *59*
Kroneckerprodukt, 50
kvadratisk form, 37, 47
kvadratisk matrix, 9
kvotienttest, 142

landmåling, 135
linearkombination, 3
lineær afbildning, *5*, *10*
lineær afhængighed, 3
lineær funktionel relation, 199
lineær ligning, løsning af, 22
lineær regressionsanalyse, 157 f
lineær uafhængighed, 3
lineært bånd, 120, 130
Little Jiffy, 383
logistisk kurve, 211
logit, 211

Mahalanobis afstand, 321
    generaliseret, 338
matrix, *8*
    regulær, 10
matrixprodukt, 9
matrixsum, 9
max R^2 improvement, 192
maximum likelihood estimation, 109
MDISC, 338 f
middelkvadratafvigelsessum
    forventet værdi af, 221, 223, 225, 233, 237, 244, 247, 248, 250, 251, 256, 257 f, 258
middelværdi, 57
Moore-Penrose invers, 28
multicollinaritet, 202
multipel korrelationskoefficient, *90*, 161

$N_p(\mu, \Sigma)$, 64
neutralt element, 2
norm, 51
normal fordeling
    flerdimensional, *64*
    todimensional, *77*
normalligninger, 114
nulrum, 7

oblimin rotation, 386
oblique rotation, 386
ortogonal matrix, 30
ortogonal regression, 96, 198 f
ortogonal transformation, 32
ortogonale polynomier, 170 f
ortogonale vektorer, 30, 51, 53
ortonormal basis, 30

partiel korrelationskoefficient, *83*, 162
positiv definit, *37*
positiv semidefinit, *37*
prediktion, 181, 202, 207
prediktionsinterval, *135*
prikprodukt, *51*
principal komponent, 349
principale komponenter, 201, 344 f
principale koordinater, 386
probability associated with largest discriminant function, 338
projektion, 5, 51
præcision, 66
pseudoinvers afbildning, *20*
pseudoinvers matrix, 1, *18*, 47, 51
pythagoræiske læresætning, 51

Q-modus, 34, 37, 383
quartimax rotation, 364

R^2, 161
R^n, 3, 4, 8
R-modus, 34, 37
rang af afbildning, 13
rang af matrix, 13, 35
Rayleigh's kvotient, 39
regression, *93*
regressionsanalyse, 157 f
    efter ortogonale polynomier, 170 f
    flerdimensional, 285, 291
    ikke lineær, 207 f
regressionsligning
    valg af bedste, 180 f
    backwards elimination, 184 f, 191
    forward selection, 186 f, 191
    samtlige regressioner, 183 f, 192
    stepwise regression, 189 f, 191
regulær matrix, 10, 13, 14
reproduktivitessætning for normal fordeling, 74
residual, 118, 163
residualplot, 163
ridge estimator, 203

ridge regression, 202 f
ridge trace, 206, 208
RSTUDENT, 169
rækkevektor, 8

SAK, 143, 296, 298
  beregning, 252 f
  frihedsgrader, 254, 255
semidefinit, 37
sideunderrum, 3
similaritetsmål, 385
similære matricer, 13
singulær værdi, 35
skalainvarians, 381
Skalarmultiplikation, 9
skalarmultiplikation, 2
skalarprodukt, *51*
spaltningssætningen, *96*
span, 3
spejling, 32
spektraldekomposition for matrix, 31
spektrum for matrix, 31
spline funktion, 212 f
spor af matrix, *46*
stepwise regression, 189 f
stokastisk matrix, 57
STUDENT RESIDUAL, 169
studentized residual, 169
støtte, 66
succesiv testning, 147
symmetrisk matrix, 9, 30
søjlevektor, 8

tabsfunktion, 309, 328
tensorprodukt, 50
test for/i
  bedste diskriminantfunktion, 323
  diagonalstruktur i dispersionsmatrix, 304
  egenværdi i dispersionsmatrix, 350
  ens dispersionsmatricer, 306
  faktormodel, 382
  flerdimensional generel lineær model, 288 f
  flerdimensional variansanalyse, 294 f
  forudsætninger I regr. analyse, 162 f
  middelværdi, 273 f, 279 f, 321
  proportional dispersion, 305
  uafhængighed, 304, 355
  yderligere information i diskr.an., 326
tr, 47
transponering, 9, 10, 51

U, 289
$U(p, q, r)$, 289
uafhængighed, 69
ukorreleret, 63
underrum, 3

varians, biologisk, 242
variansanalyse, 215 f, 294 f
  flere faktorer, 251 f
  hierarkisk klassifikation, 227 f, 232 f, 246 f, 255 f
  krydsklassifikation, 227 f, 237 f, 243 f, 252 f, 261
  systematisk model, 231 f
  tilfældig model, 216, 218, 222 f, 230 f
  tosidet, 227 f
  tosidet flerdimensional, 296 f
  trefaktor, 242 f
  vekselvirkning, 241, 252 f
variansanalyse fortsat
  robusthed, 260 f
variansanalyse,systematisk model, 219 f
variation
  inden for grupper, 225, 295, 334
  mellem grupper, 225, 295, 334
  mellem hovedgrupper, 232
  mellem undergrupper i.f. hvd. gr., 232
  spaltning af totale, 173, 225, 232, 237, 244, 245, 247, 248, 250, 251, 254, 295, 298
  systematisk, 216
  tilfældig, 216
variationsanalyse
  systematisk model, 216
  vekselvirkning, 237
variationsbredde, 165

varimax rotation, 364, 386
VC, 63
vektoraddition, 2
vektorrum, *2*
vinkel, 53
vægtet regression, 158

$W(n, \Sigma)$, 102
Wilk's Λ, 289, 335
Wishart fordeling, *102*

Appendiks A
Det græske alfabet

	Bogstavnavn	*Udtale*	*Gengivelse*
A α	alfa	[a][a:]	a
B β	bēta	[b]	b
Γ γ	gamma	[g]	g
Δ δ	delta	[d]	d
E ϵ	epsilon	[e]	e
Z ζ	zēta	[ts,s]	z
H η	ēta	[æ:]	ē
Θ ϑ θ	thēta	[θ,th,t]	th,t
I ι	iōta	[i][i:]	i
K κ	kappa	[k]	k
Λ λ	lambda	[l]	l
M μ	my	[m]	m
N ν	ny	[n]	n
Ξ ξ	ksi	[ks]	ks(x)
O o	omikron	[o]	o
Π π	pi	[p]	p
P ρ	ro	[r]	r
Σ σ ς	sigma	[s]	s
T τ	tau	[t]	t
Y υ	ypsilon	[y][y:]	y
Φ ϕ	fi	[f]	f(ph)
X χ	khi	[x,ç,kh,k]	ch(kh)
Ψ ψ	psi	[ps]	ps
Ω ω	ōmega	[å:]	ō