MønsterGenkendelse

Forår 1996

S. I. Olsen

Dette skrift er 2. udkast til et notesæt til brug i kurset Mønstergenkendelse. Noterne dækker primært områderne:

- Statistiske mønstergenkendelse,
- Lineær regression,

Desuden er emneområder som *Estimation* og *Neurale net* behandlet i et vist omfang. Noterne dækker derimod **ikke** områderne: Syntaktisk mønstergenkendelse, Informationsteori, Kodning, Frekvensanalyse, herunder fouriertransformationen, mv.

Formålet med noterne er at give læseren en introduktion til de behandlede emner. Noterne kan således på ingen måde tilfredsstille den mere erfarne læser. Noterne omfatter emner, der traditionelt er henføres til områderne *Matematik*, *Statistik*, *Numerisk analyse*, *Optimering* mv. Læsere med solidt kendskab til disse områder vil nok finde noterne mangelfulde. Formålet med noterne er da også at kombinere basal viden fra disse områder på en forhåbentlig brugbar måde, mere end at give en udtømmende beskrivelse af de teknikker, der er udviklet inden for de nævnte områder.

Da notesættet er under udvikling, må der forventes at forekomme en del fejl i disse, såvel som emner, der kun er behandlet kursorisk eller på anden måde utilfredsstillende.

Noterne henvender sig til datalogistuderende med bestået førsteårskursus i matematik. Det forudsættes således at læseren har et modent (operationelt) forhold til matematik. For at lette læsningen og forståelsen af notationen er noterne forsynet med opsummerende appendix om basal lineær algebra og sandsynlighedsregning samt en ultrakort introduktion til nogle optimeringsmetoder. Disse afsnit er kun tænkt som en hjælp, og kan på ingen måde erstatte egenligt undervisningsmateriale inden for de nævnte emner.

Til hver kapitel i noterne er der knyttet et antal opgaver. Visse af disse vil være protopyper på opgaver der kan forekomme i den skriftlige eksamen. Andre af opgaverne har til formål at træne læseren i praktisk anvendelse af stoffet, eller at uddybe delområder, som kun bliver omtalt perifert i noterne.

Der eksisterer en lang række lærebøger inden for de fagområder som noterne dækker. Derimod er det vanskeligt at finde én bog, der dækker hele stofområdet (hvilket begrunder behovet for noterne). Et eksempel på en god bog, der dækker forholdsvist bredt er [1]. Referencen kan findes sidst i noterne i afsnittet "Litteratur".

Om formålet med kurset

I formålsbeskrivelsen for kurset nævnes:

> Formålet med kurset er at tilegne de studerende en basal viden om grundlæggende algoritmer til **analyse**, **modellering**, og **fortolkning** af data. Centralt er kendskab til metoder hvorved **mønstre** i data kan **genkendes**, samt hvorledes data kan **klassificeres** i forhold til et sæt af **prototyper**. Ved genkendelse af mønstre kan forekomster af disse repræsenteres ved en reference til en prototype. Herved kan en (evt. hierakisk) **struktur**, der afspejler den **syntaktiske** sammensætning af data opbygges. Væsentlige anvendelser er **kodning** og **kompression** af data. Ofte er data genereret af en process hvis struktur og egenskaber kun er delvist kendt. Mønstergenkendelse kan i denne sammenhæng forstås som opgaven at afdække processens struktur og egenskaber ud fra de observerede data, dvs. at modellere data. Ofte svarer de observerede data ikke eksakt til modellens forudsigelser, men er behæftet med støj og irrelevante data. I disse tilfælde er en **statistisk** analyse nødvendig. Væsentlige modeller er de **lineære systemer**, og **fit** af disse til data. Ved modeller der beskriver frekvenssammensætningen af data er **fourieranalysen** central. I **samplingsætningen** redegøres for hvilken information, som kan udtrækkes af et frekvensbegrænset kontinuert signal ved at tage stikprøver.
>
> Det er ikke hensigten med kurset at uddanne eksperter i mønstergenkendelse, eller at fokusere på specifikke anvendelser. Derimod er det hensigten at give en introduktion til en række af generelt anvendelige metoder, og at fokusere på nogle klassiske og bredt benyttede algoritmer.

Den del af kurset som disse noter beskriver omhandler statistisk dataanalyse, lineær regression, og klassifikation, herunder neurale net. Det er håbet at noterne vil give den datalogisk relevante og nødvendige baggrund ved behandling af data (tal), samt at være til nytte i enhver situation hvor data skal analyseres og vurderes. Nogle eksempler på spørgsmal, som det er håbet at noterne vil give svar på er:

- Hvordan beregner man middelværdien af en række indtastede tal hvor i blandt der forekommer indtastningsfejl, og hvordan måler man hvor sikker en middelværdibestemmelse er.
- Hvordan bestemmer man koefficienterne i det n'te grads polynomium, der passer bedst til et sæt af målinger.
- Hvodan karakteriserer man en prototype ud fra en population af stikprøver, og hvordan klassificerer man en måling som stammende fra én af et antal klasser af prototyper.
- Hvilken strategi skal man vælge når man skal miminere den forventede fejl, eller maksimere det forventede udbytte.
- Hvor stor en klasse af funktioner kan et neuralt netværk modellere.

Notation

Nedenstående afsnit opsummerer de konventioner for notation, som jeg har forsøgt at følge gennem noterne.

$\mathcal{R}$	Kursiverede store bogstaver betegner mængder.
$\mathbf{X}$	Fede store bogstaver betegner matricer.
$\mathbf{x}$	Fede små bogstaver betegner vektorer
$\bar{\mathbf{x}}$	Overstreget fede små bogstaver betegner middelværdi af en vektor.
$\hat{\mathbf{x}}$	Hattede fede små bogstaver betegner et estimat.
$E(\cdot)$	Operatoren E betegner "forventet værdi af".
$p(\cdot)$	Operatoren p betegner "sandsyligheden for".
$p(A\|B)$	Betinget sandsyligheden for A givet B.
$\mathbf{x}_{\cdot j}$	Søjlevektor af matrice.
$\mathbf{x}_{i\cdot}$	Rækkevektor af matrice.
$\mathbf{x}^t$	Vektortransponering.
(x_{ij})	Matrice af elementer x_{ij} i række i og søjle j.
$f(x)$	Funktion af reelt argument.
$f[x]$	Funktion af heltalligt argument.

Der vil utvivlsomt være en lang række symboler, konventioner, og notationer, der ikke er kommet med i ovenstående liste. Ligeledes vil der formodentlig være tvetydigheder, der bør afklares. Som læser vil du give forfatteren en uvurderlig hjælp, hvis du vil videregive ham observationer om sådanne tilfælde.

Forelæsningsplan

Noterne dækker 10 enkeltforelæsninger. Det er planen at gennemgå stoffet på 5 uger, som nærmere meddelt ved kursusstarten. Indholdet af de 10 enkeltforelæsninger er kort skitseret nedenfor.

F 1 **Grundlæggende statistik**
Stokastisk variabel, tæthedsfordeling, middelværdi, varians, estimation af middelværdi og varians, momenter, empiriske fordelinger, robuste estimatorer.

F 2 **Grundlæggende lineær algebra**
Kovarians, vektorfunktioner, kovariansmatricer, auto-korrelationsmatricer, egenværdier og egenvektorer,

F 3 **Estimation**
Test af middelværdi og varians, betingede sandsynligheder, Bayes regel, maksimum-likelihood estimation, Bayes estimation.

F 4 **Featureselektion**
Principalkomponentanalyse

F 5 **Lineær regression**
Lineære modeller, løsning af kvadratiske og overbestemte lineære ligningssystemer, normalligninger, den pseudo-inverse matrice, designmatrice, residualer, mindste kvadraters metode.

F 6 **Outliers og mindste kvadraters metode**
Outliers, mindste kvadraters metode i data med outliers.

F 7 **Klyngeanalyse**
Minimalt udspændende træer, k-d-træer, K-means algoritmen, maximin-algoritmen.

F 8 **Statistisk klassifikation**
Klassifikation, Beslutningsfunktioner, Minimal afstands klassifikatorer.

F 9 **Diskriminantanalyse og Neurale net**
Statistisk klassifikation, Beslutningsfunktion for normale mønstre. Typer af kunstige neurale net, feed-forward neurale net som implementation af lineære beslutningsfunktioner.

F 10 **Neurale net**
Generaliseringsevne, oplæring af neurale net, back propagation algoritmen.

Indholdsfortegnelse

Kapitel 1

Statistisk Dataanalyse

Ved analyse af data, med det formål at ekstrahere mønstre i disse, er det vigtigt (inden fastlæggelse af en bestemt analysemetode), at gøre sig klart hvilken grad af regelbundenhed contra tilfældighed som de aktuelle data udviser. Data kan groft rubriceres som:

- Stærkt regelbundne.
- Svagt regelbundne.
- Tilfældige.

Hvis data er fuldstændigt regelbundne kan strukturen i data ofte analyseres bedst ved at opstille en grammatik for de regler som definerer hvorledes de enkelte dataelementer kan forekomme. Teknikker til analyse af fuldstændig regelbundne data er veludviklede inden for *programanalyse*, hvor syntaksanalysatorer kan konstrueres automatisk ud fra de opstillede regler. Denne type af dataanalyse er en grunddiciplin inden for datalogi, og er grundigt behandlet på andre kurser. Vi skal derfor ikke beskrive disse metoder yderligere.

Hvis data er forholdsvist stærkt regelbundne, men reglerne kun er delvist kendte, er metoder fra emneområdet *kunstig intilligens* ofte anvendelige. Mange af disse metoder går under kælenavnet *ekspertsystemer*, fordi de fundne regler ofte er fremkommet ved formalisering af eksperters viden om regelbundetheden. I forhold til grammatikker for programmeringssprog, hvor reglerne ofte er hierakisk opbygget, er de regler, der benyttes i ekspertsystemer ofte (men ikke altid) sideordnet. Viden i et ekspertsystem formaliseres ofte ved en mængde af såkaldte *produktionsregler*. En produktionsregel udtrykker at hvis en række betingelser er opfyldt, da er det muligt at drage en (del)konklussion. Ofte er denne tilknyttet et mål for troværdigheden af slutningen. Dataanalysen foretages af en *inferensmaskine*, der givet data og sættet af produktionsregler iterativt anvender en produktionsregel. Gradvist opbygges et sæt af *hypoteser* om de aktuelle data. Til hver hypotese er knyttet et troværdighedsmål. Når en hypotese bliver for utroværdig slettes denne af listen af hypoteser. Når det ikke længere er muligt at forbedre hypoteserne (ud fra de aktuelle data og de kendte regler) uvælges den (de) bedste hypoteser. Ovenstående meget kortfattede beskrivelse af funktionsmåden af et ekspertsystem er naturligvis på ingen måde fyldestgørende, og udelukkende tænkt som et holdepunkt i forhold til hvilken statistiske analysemetoder kan forstås. En dybere forståelse af de mange metoder inden for kunstig intelligens ligger uden for rammerne af dette kursus.

Hvis data indeholder støj, dvs. et element af tilfældighed, er det ofte nødvendigt at ty til statisk funderede metoder for dataanalyse. Det er typisk at datamaterialet er forholdsvist stort, hvorimod de modeller der anvendes er (matematisk set) forholdsvist enkle. Man skelner mellem *parametriske* og *ikke-parametriske* modeller. I en parametrisk model af data (f.eks. en lineær sammenhæng mellem en række værdier) ønsker man at *estimere* parametrene, dvs. at finde den model der *fitter* data bedst muligt. Forskellen, fejlen, mellem de observerede data og modellens forudsigelse, ønskes med andre ord så lille som muligt. Estimationen kan beskrives ved at bestemme de mest sandsynlige parametre, som kan forklare/beskrive de observerede data. I ikke-parametriske modeller er målet for analysen at beskrive de statistiske egenskaber ved data, herunder middelværdi, varians etc. Sådanne analyser er hensigtsmæssige, når den eneste regelbundethed af data er fastlagt ved fordelingsfunktionen for dataværdierne. I grænseområdet mellem parametriske og ikke-parametriske analyser findes det område, der går under navnet *kaosteori*. Kaotiske data er ekstremt svagt regelbundne.

I dette kapitel introduceres til de statisktisk baserede analyse- og estimationsmetoder. I de efterfølgende kapitler konkretiseres en række analysemetoder, der alle er stærkt funderet i statistikken. Årsagen til, at vi har valgt at lægge så stor vægt på statistisk dataanalyse er, at disse metoder er generelt anvendelige, meget udbredte, og solidt teoretisk funderet. Grundigt kendskab til disse metoder er ofte en nødvendighed ved studier af metoder inden for andre områder, f.eks. kunstig intilligens eller neurale net. I Appendix C er en række basale definitioner inden for sandsyligheder og statistik kort gennemgået. Nedenstående gennemgås nogle grundlæggende metoder. Appendix C indeholder en række supplerende metoder. Nedenstående gennemgang er rettet mod de anvendelser, der senere beskrives i noterne, og kan på ingen måde erstatte egenligt undervisningsmateriale inden for hverken statistik eller sandsynlighedsregning.

1.1 Stokastisk variabel

En stokastisk variabel x defineret på udfaldsrummet Ω, og med fordelingsfunktion F, er en størrelse hvis værdi ikke kendes eksakt. Eksempelvis vides ikke eksakt hvor mange biler der mellem kl. 9:00 og 10:00 kører af Jagtvej. Det er muligt at estimere værdien af den stokastiske variabel ved at foretage målinger. Her vil man kunne iagttage at næppe to målinger vil være ens. Ved at foretage mange målinger, og ved at tage gennemsnit af disse vil man kunne estimere middelværdien af den stokastiske variabel. Dette er imidlertid en meget grov karakterisering. En fuldstændig karakterisation af en stokastiske variabel er bestemt ved fordelingsfunktionen F for variablen. Fordelingsfunktionen $F(x)$ vil for ovenstående eksempel udtrykke sandsynligheden for at antallet af optalte biler er mindre end x. I dette tilfælde er udfaldsrummet diskret (der kan være 0, 1, 2, ...) biler, men ikke

1.37 bil. I andre stituationer vil udfaldsrummet være en delmængde af $\mathcal{R}^n$ (eksempelvis mængen af fobrændt benzin ved kørsel på Jagtvej). For $n = 1$ er $F(x)$ en reel funktion af en variabel. Den stokastiske variabel kaldel kontinuert, hvis $F(x)$ er kontinuert og differentiabel. Tæthedsfunktionen $f(x)$ er da defineret som differentialkvotienten af $F(x)$.

1.2 Tæthedsfordeling

I dette afsnit gives eksempler på nogle få hyppigt anvendte tæthedsfunktioner for stokastiske variable, og det beskrives hvorledes tæthedsfunktionen ændrer sig ved simple transformationer af den stokastiske variabel.

Antag at den stokastiske variabel er defineret på et interval $[a, b]$. En *ligefordeling* (eller uniform fordeling) er da givet ved:

$$F(x) = \frac{x-a}{b-a}$$

Her er sandsynlighedsmassen altså spredt ligeligt ud over definitionsmængden. Diamentralt modsat kan hele sandsynlighedsmasses være samlet i ét punkt a.

$$F(x) = \delta(x-a) = \begin{cases} 1 & \text{hvis } x = a \\ 0 & \text{ellers} \end{cases}$$

Funktionen δ kaldes *Diracs deltafunktion.*

Der findes et væld af fordelinger, der ofte er begrundet i en fysisk model. *Eksponentialfordelingen* er (for $x \geq 0$) bestemt ved:

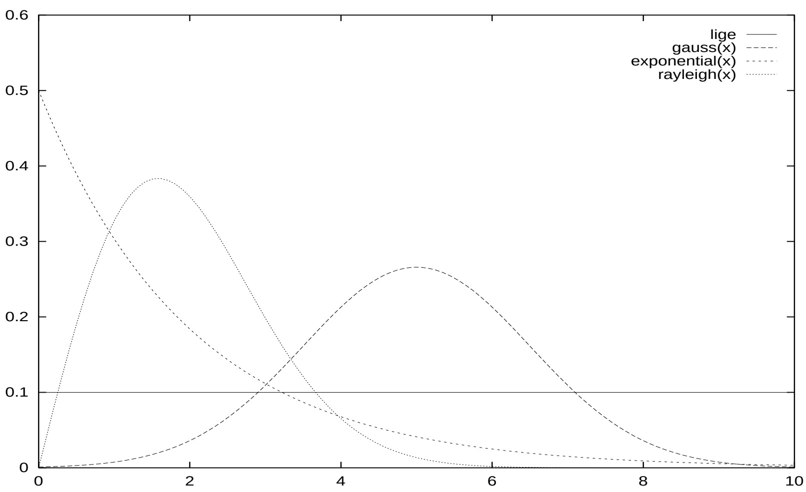
$$F_{eks}(x) = 1 - e^{-\lambda x}$$

svarende til tæthedsfunktionen $f(x) = \lambda e^{-\lambda x}$. Det klassiske eksempel på en eksponentialfordeling opstår ved betragtning af en mængde af agenter, der uafhængigt af hinanden og uafhængigt af agenternes alder, spontant beslutter at udsende en meddelelse. Da er fordelingsfunktionen for antallet af udsendte meddelelser givet ved $F_{eks}(x)$. Radioaktive henfald modelleres godt ved en eksponentialfordeling. Parameteren λ bestemmer hvor hurtigt $f(x)$ aftager.

Normalfordelingen eller *Gauss fordelingen* er givet ved:

$$f(x) = G(\mu, \sigma)(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Normalfordelingen er interessant på grund af en lang række egenskaber. Antag at den stokastiske variabel x fremkommer ved summation af uendelig mange meget små led (stokastiske variable), der alle har samme fordeling f_0. Da er x normalfordelt uafhængig af fordelingen f_0. Dette resultat kaldes *Den centrale grænseværdisætning* . Støj modelleres ofte ved en normalfordeling. Vi skal senere i noterne intensivt benytte en sådan antagelse.



Figur 1.1: Illustration af en ligefordeling på intervallet [0,10], en normalfordeling med $\mu = 5, \sigma = 1.5$, en eksponentialfordeling med parameteren $\lambda = 0.5$, samt en rayleigh fordeling $2\lambda x e^{-\lambda x^2}$ med parameteren $\lambda = 0.2$.

Normalfordelingen har to parametre μ og σ. Disse kaldes *middelværdien* hhv. *spredningen* af fordelingen. Disse navne benyttes generelt til karakterisering af enhver fordeling, men har altså en særlig betydning for normalfordelingen.

Andre fordelinger er konstrueret på basis af simplere fordelinger ved transformationer af stokastiske variable. En vigtig sådan fordeling er gamma-fordelingen. Det kan vises at variansestimatet for en følge af normalfordelte stokastiske variable (med samme middelværdi) er gamma-fordelt. Snævert relateret er χ^2-fordelingen $Q(\chi^2, \nu)$, der angiver sandsynligheden for at en sum af kvadrater af ν normalfordelte stokastiske variable, med varians lig 1, er større end χ^2. Størrelsen ν kaldes antallet af frihedsgrader. Vi skal senere benytte denne fordeling ved test af hvorvidt to fordelinger er ens. Der findes en række fordelinger, der benyttes i test af forskellige hypoteser, eksempelvis Students t-fordeling, F-fordelingen, binomial-fordelingen, Poisson-fordelingen, beta-fordelingen, etc. Selv om kendskab til disse fordelinger er nødvendig for en dybere forståelse af mange statistiske metoder, er de ikke essentielle for dette kursus.

I mange sammenhænge kan det være nyttigt at beskrive de observerede data y som en transformation ϕ af en stokastisk variabel x. Funktionen ϕ modellerer ofte et fysisk system. Hvis funktionen ϕ er kendt er det i visse tilfælde muligt at beregne den teoretiske fordeling af observationerne $y = \phi(x)$, hvor y her betragtes som en stokastisk variabel. Antag at x er en kontinuert reel stokastisk variabel med tæthed f, defineret på intervallet [a,b], og antag at $\phi :]a, b[\rightarrow]c, d[$ er bijektiv. Antag yderligere at $\psi = \phi^{-1}$ eksisterer og er kontinuert. Da er tæthedsfunktionen g for den transformerede stokastiske variabel y bestemt ved:

$$g(y) \;=\; f(\psi(y))\,|\psi'(y)| \tag{1.1}$$

hvor $y \in [c,d]$, og $\psi' = \frac{\partial \psi}{\partial y}$. Det er muligt at generalisere sætningen til tilfældet hvor ϕ er stykvis bijektiv med kontinuert differentialkvotient. Der er to oplagte anvendelser af sætningen. Hvis transformationen er kendt, er det muligt at teste (se senere) hvorvidt antagelsen om fordelingen f er holdbar. Hvis fordelingen f er kendt, men transformationen ϕ er parametriseret, udgør ligningen et grundlag for estimation af disse parametre ud fra den observerede tæthedsfunktion g.

Eksempel
Hvis x eksempelvis er uniformt fordelt på intervallet [0,1], dvs. at $f(x) = 1$, og $\phi(x) = x^2$, da fås at $\psi(y) = \sqrt{y}$, at $\psi'(y) = 1/(2\sqrt{y})$, og dermed at $g(y) = f(\sqrt{y})\ |1/(2\sqrt{y})| = \sqrt{y}/(2y)$.

Eksempel slut

En tredie anvendelse af ovenstående resultat er, at normalisere et sæt data med fordeling f til et nyt sæt data, der har fordelingen g. Denne anvendelse foretages ofte med billeder, hvor de fleste billedelementer har næsten ens værdi, dvs. at sandsynlighedsmassen er koncentreret på en lille del af intervallet [a,b]. I dette tilfælde kan det være vanskeligt at skelne detaljer i billedet. Før transformationen normaliseres alle værdier til intervallet [0,1] ved division med den maksimale intensitet. Det kan let vises at hvis $\phi(x)$ vælges som fordelingsfunktionen for x, dvs.

$$\phi(x) \;=\; \int_0^x f(w)dw$$

da vil g (i det kontinuerte tilfælde) være konstant lig 1 på intervallet [0,1]. Efter transformationen multipliceres de transformerede værdier med den maksimale intensitet, og resultatet trunkeres til nærmeste mulige intensitetsværdi (ofte heltallig). Effekten af transformationen er at alle intensiteter udnyttes ligeligt (i det kontinuerte tilfælde). Dette vil ofte bringe detaljer, der før var uskelnelige, klart frem. Bemærk at der i det diskrete tilfælde vil gælde at antallet af forskellige intensiteter efter transformationen højst vil være lig antallet af forskellige intensiteter før transformationen.

1.3 Middelværdi og varians

Den statisktisk *forventede værdi*, også kaldt *middelværdien*, af en reel stokastisk variabel x med tæthedsfunktion f, er bestemt ved:

$$E(x) \;=\; \mu \;=\; \int_{\mathcal{R}} t \cdot f(t)dt \tag{1.2}$$

Hvis integralet ikke er konvergent tillægges x ingen middelværdi. Generelt gælder om en funktion $g : \mathcal{R} \rightarrow \mathcal{R}$ at:

$$E(g(x)) \;=\; \int_{\mathcal{R}} g(t)f(t)dt \tag{1.3}$$

Betragtes transformationen $z = z(x,y) = ax + by$, hvor x, y, og z er stokastiske variable og a og b er reelle tal, da er $E(z) = aE(x) + bE(y)$. Middelværdioperatorer er således lineær.

Variansen af en stokastisk variabel med middelværdi μ betegnes $\sigma^2 = \sigma^2(x) = Var(x)$ og er defineret ved:

$$\sigma^2 \;=\; E(|x-\mu|^2) \tag{1.4}$$

Hvis den stokastiske variabel $|x-\mu|^2$ ingen middelværdi har, sættes $\sigma^2 := \infty$. Den ikke negative kvadratrod af variansen kaldes *spredningen* eller *standard afvigelsen* og betegnes $\sigma = \sigma(x)$. Om variansen gælder der at:

$$\begin{array}{ll} 1 & \sigma^2(x) \geq 0 \\ 2 & \sigma^2(x) = E(x^2) - [E(x)]^2 \\ 3 & \sigma^2(ax) = a^2\sigma^2(x) \quad \text{for } a \in \mathcal{R} \\ 4 & \sigma^2(x+a) = \sigma^2(x) \quad \text{for } a \in \mathcal{R} \end{array}$$

Kravet til beregning af middelværdi og varians/spredning for en stokastisk variabel x er altså, at tæthedsfunktionen for x er kendt. Hvis dette ikke er tilfældet er det muligt at *estimere* middelværdien og variansen på basis af et antal stikprøver (samples) af x. Det antages at stikprøverne er uafhængige. Estimaterne vil være usikre hvis antallet af stikprøver er lille og vil blive mere nøjagtige jo flere stikprøver, der er til rådighed. Basis for estimationen er at erstatte den statistisk forventede værdi med et gennemsnit. For n stikprøver fås:

$$\hat{\mu} \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1.5}$$

$$\hat{\sigma}^2 \;=\; \frac{1}{n}\sum_{i=1}^{n} x_i^2 \;-\; [\frac{1}{n}\sum_{i=1}^{n} x_i]^2 \tag{1.6}$$

hvor x_i er den i'te stikprøve af x.

Da estimatet $\hat{\mu}$ er en summation af stokastiske variable, er det selv en stokastisk variabel, og har, som sådan, en middelværdi og en varians. Middelværdien af $\hat{\mu}$ er:

$$E(\hat{\mu}) \;=\; E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \;=\; \frac{1}{n}\sum_{i=1}^{n} E(x_i) \;=\; \frac{1}{n}\sum_{i=1}^{n} \mu \;=\; \mu \tag{1.7}$$

Den forventede værdi af estimatet er altså lig middelværdien selv. Et estimat, der har denne egenskab, kaldes et *unbiased estimat* . For variansen af $\hat{\mu}$ fås:

$$\begin{aligned} Var(\hat{\mu}) \;&=\; E([\hat{\mu}-\mu]^2) \\ &=\; E([\frac{1}{n}\sum_{i=1}^{n} x_i \;-\; \mu]^2) \\ &=\; \frac{1}{n^2}E([(x_1-\mu)+(x_2-\mu)+\cdots+(x_n-\mu)][(x_1-\mu)+(x_2-\mu)+\cdots+(x_n-\mu)]) \\ &=\; \frac{1}{n^2}E(\sum_{i=1}^{n}(x_i-\mu)^2) \;+\; \sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} E((x_i-\mu)(x_j-\mu)) \\ &=\; \frac{1}{n^2}E(\sum_{i=1}^{n}(x_i-\mu)^2) \\ &=\; \frac{1}{n}\sigma^2 \end{aligned} \tag{1.8}$$

hvor vi har udnyttet at de enkelte stikprøver er uafhængige, hvorved $E([x_i - \mu][x_j - \mu]) = E(x_i - \mu)E(x_j - \mu) = 0$ for $i \neq j$. Variansen af estimatet er altså $1/n$ gange variansen af x. Heraf ses at usikkerheden (målt ved variansen) af $\hat{\mu}$ går mod 0 når $n \to \infty$. En estimator, der har denne egenskab kaldes en *konsistent estimator*.

For variansen er en vurdering af usikkerheden på estimationen lidt mere kompliceret. Problemet er, at forkundskab til middelværdien er nødvendig for beregning af variansen, jvf. definitionen (1.4). For en mængde af n stikprøver, hvorom vi intet ved, siges antallet af *frihedsgrader* at være lig n. Hver gang vi bestemmer en parameter i tæthedsfunktionen øves vores viden. Tilsvarende mindskes friheden blandt de observerede data. Antallet af frihedsgrader reduceres med én. Efter bestemmelsen af middelværdien er antallet af frihedsgrader derfor $n - 1$. Det kan vises at estimationen (1.6) er biased, hvorimod estimationen:

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}[x_i - \mu]^2 \tag{1.9}$$

er unbiased. Det ses her, at der divideres med antallet af frihedsgrader, og ikke med antallet af stikprøver. I praktisk anvendelse bør n være så stor, at forskellen mellem de to estimater er forsvindende.

I mange situationer er der grund til at tro, at stiprøverne stammer fra en kendt fordeling. I dette tilfælde er opgaven derfor at estimere parametrene i fordelingen. Er der eksempelvis grund til at tro, at fordelingen er normal, da er denne fuldstændigt specificeret ved middelværdien og variansen. Spørgsmålet, om den empiriske fordeling (repræsenteret ved stikprøverne) faktisk stemmer overens med den estimerede fordeling diskuteres i et følgende afsnit.

Hvis der ikke er grund til at tro, at stikprøverne stammer fra en bestemt fordeling er problemet at karakterisere den empiriske fordeling. Middelværdien og variansen er her to meget beskrivende størrelser. Mange andre karakteriseringer kan imidlertid være nyttige. Det p'te *centrale moment* m_p er defineret ved:

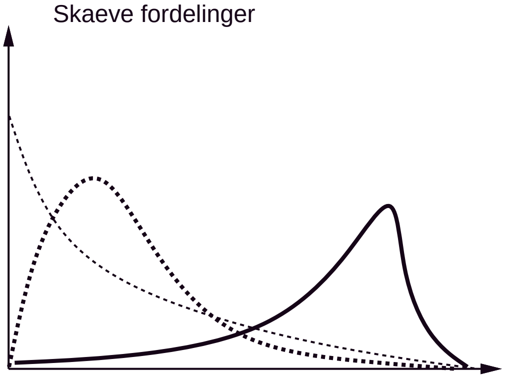
$$m_p = E([x - \mu]^p) \tag{1.10}$$

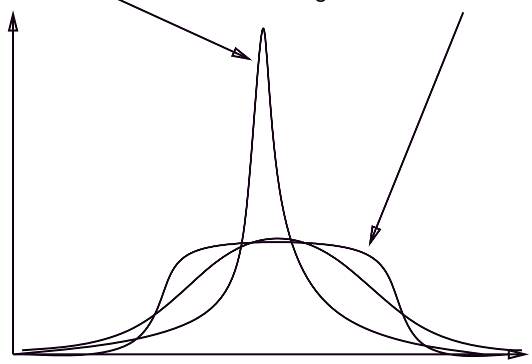
Det første centrale moment er lig 0, det andet centrale moment er lig variansen. To ofte benyttede karakteriseringer, *skævhed* og *kurtiosis* af en fordeling er defineret ud fra de centrale momenter af 3. og 4. orden:

$$Skew(x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{x_i - \mu}{\sigma}\right]^3 = \frac{m_3(x)}{m_2(x)^{3/2}} \tag{1.11}$$

$$Kurt(x_1, \ldots, x_n) = \left\{\frac{1}{n}\sum_{i=1}^{n}\left[\frac{x_i - \mu}{\sigma}\right]^4\right\} - 3 = \frac{m_4(x)}{m_2(x)^2} - 3 \tag{1.12}$$

I modsætning til middelværdien og spredningen (der har samme enhed som observationerne selv) er skævheden og kurtiositeten dimmensionsløse. Skævheden karakteriserer graden af asymmetri af tæthedsfunktionen (omkring middelværdien). Kurtiosis målet karakteriserer hvor flad contra spids tæthedsfunktionen er. Årsagen til subtraktion af konstanten 3 i målet for kurtiositet er, at målet hermed vil give værdien 0 for en normalfordeling. Hvis målet er negativt vil fordelingen være fladere end en normalfordeling, hvis målet er positivt vil den empiriske fordeling være spidsere end en normalfordeling. Det skal bemærkes at momenter af orden højere end 2 (pga. potensopløftningerne) er meget følsomme over for variationer i halerne af en empirisk fordeling.





Middelværdi samt mål baseret på de centrale momenter er ikke de eneste interessante karakteriseringer af en fordeling. F.eks. er positionen af maksimum (toppunktet) for den empiriske tæthedsfunktion (eng. mode) ofte nyttig. *Medianværdien* angiver den midterste værdi i en sorteret følge af målinger. I det kontinuerte tilfælde gælder at medianen $= F^{-1}(0.5)$. Den øvre hhv. nedre α-kvartil angiver værdien hvor α % af stikprøverne er mindre hhv. større end kvartilen.

Om empiriske fordelinger gælder generelt, at de sjældent er så pæne som man kunne ønske. Dette gælder især hvis den proces, der har genereret de observerede data skifter mellem flere, iøvrigt nogenlunde stabile, tilstande. I dette tilfælde vil den empiriske tæthedsfunktion ofte have flere markant forskellige toppe. En fordeling kaldes *unimodal* hhv. *bimodal* hhv. *multimodal* hvis den har én hhv. to hhv. mange sådanne toppe. Hvis en fordeling er bimodal, vil en estimation af middelværdi og varians af hele fordelingen ikke give mening. Inden estimation af disse parametre er det derfor tilrådeligt at checke fordelingens modalitet. Hvis man ønsker at teste om en empirisk fordeling passer godt med en teoretisk fordeling, kan *goodness-of-fit-test* metoden, der beskrives i et følgende afsnit af noterne, anvendes. Der findes en lang række algoritmer til analyse af modaliteten af en empirisk fordeling, såvel som metoder til seperation af fordelingen i plausible komponenter. Disse metoder er ofte baseret på ad'hoc kriterier. Det vil føre for vidt her et beskrive sådanne metoder.

1.4 Robuste estimater

Som beskrevet i større detalje senere i noterne er de sædvanlige estimater af såvel middelværdien som variansen meget følsomme over for eksistensen af stikprøver, der ligger i en af halerne af fordelingen. Blot én stikprøve er tilstrækkeligt afvigende kan et vilkårligt estimat fremkomme. Estimater, der har denne (lidt uheldige) egenskab er ikke *robuste*. Det er muligt at definere estimatorer, der kan vises at være mere robuste.

Betragt en mængde af n tal. Det sædvanlige estimat af den forventede værdi er lig gennemsnittet af tallene. Denne metode siges at have et *nedbrudspunkt* på $1/n$ fordi et vilkårligt resultat kan frembriges blot ét blandt de n tal er tilstrækkeligt afvigende. Nedbrudspunktet for en estimator beskrives i større detalje senere i noterne. Et mere robust estimat af den forventede værdi er *medianværdien* , dvs. den midterste af de sorterede n værdier. Denne estimator har et nedbrudspunkt på 0.5, fordi en erstatning af under halvdelen af værdierne med vilkårlige andre værdier ikke kan flytte medianværdien vilkårligt.

En anden robust estimator af middelværdien fremkommer ved at sortere de n tal, og beregne middelværdien af de $n(1-2\alpha)$ midterste værdier. Lad $m = \lfloor n\alpha \rfloor$. Estimatoren er da:

$$T_\alpha = \frac{1}{n-2m} \sum_{i=m+1}^{n-m} r_i \tag{1.13}$$

Metoden, der kaldes den *α-trimmede middelværdi*, kan vises at have et nedbrudspunkt på α. For $\alpha = 0$ fås det sædvanlige gennemsnit, for $\alpha = 1/2$ fås medianværdien.

En ulempe ved anvendelsen af robuste metoder er, at disse ikke er nøjagtige, eller konsistente (dette er hvad man betaler for robustheden). Hvis de n tal eksempelvis alle er heltallige vil medianværdien også være heltallig, selv om den statistisk forventede værdi er et reelt tal. Hvis fordelingen af de n tal er meget skæv (asymmetrisk) kan både medianværdien, den sædvanlige middelværdi, og den α-trimmede middelværdi ligge langt fra middelværdien, selv om ingen af de n tal er outliers. Outliers diskuteres i detaljer senere i noterne.

Et robust (men ikke nødvendigvis nøjagtigt) estimat af spredningen er *mean absolute deviation* eller MeanAD-estimatet :

$$MeanAD = \sqrt{\pi/2}\,\frac{1}{n} \sum_{i=1}^{n} |x_i - \mu| \tag{1.14}$$

Et andet robust mål af spredningen, kaldt MedianAD-estimatet (*median absolute deviation*) er:

$$MedianAD = 1.4826\; med_i \left\{ |x_i - med_j x_j| \right\} \tag{1.15}$$

Begrundelserne for konstanterne $\sqrt{\pi/2}$ og 1.4826, er at estimatorerne uden disse konstanter ville give et systematiske forkert resultat hvis de n stokastiske variable var normalfordelte. Faktoren $1.4826 = 1/\Theta^{-1}(0.75)$ kompenserer således for anvendelsen af mediafilteret i normalfordelt støj. $\Theta(x)$ er lig fordelingsfunktion for en normalfordelingen.

Da det sædvanlige skævhedsmål også behandler alle målinger ens, er dette ikke robust. Et (måske mere intuitivt) robust mål for skævheden af en fordeling er givet ved den relative afstand mellem medianværdien og 25%-kvartilen hhv. 75% kvartilen:

$$SKEW_{kvartil} = \frac{75\%kvartil - median}{median - 25\%kvartil} - 1 \tag{1.16}$$

Eksempel
Antag at vi har foretaget 49 målinger af en heltallig (ikke negativ) stokastisk variabel. Antallet af observationer er for hver værdi givet i nedenstående skema:

værdi	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
antal	1	4	9	11	8	4	3	2	2	1	1	0	1	0	0	1	0	0	1

Ved beregning af de ovenfor beskrevne mål fås:

mål	Middel	median	0.2-trimmet	25% kvartil	75% kvartil	maksimum
værdi	4.45	3	3.65	2	8	3

mål	σ	$MeanAD$	$MedianAD$	$Skew$	$SKEW_{kvartil}$	$Kurt$
værdi	3.51	3.08	1.48	1.92	4	4.02

Som det ses er fordelingen skæv således at hoveddelen af sandsynlighedsmassen ligger til venstre for middelværdien. Fordelingen er spidsere end normalfordelingen. På grund af den lange hale af fordelingen er spredningen større end de to mere robuste mål MeanAD of MedianAD. Læseren opfordres til at efterregne de angivne mål.

Eksempel slut

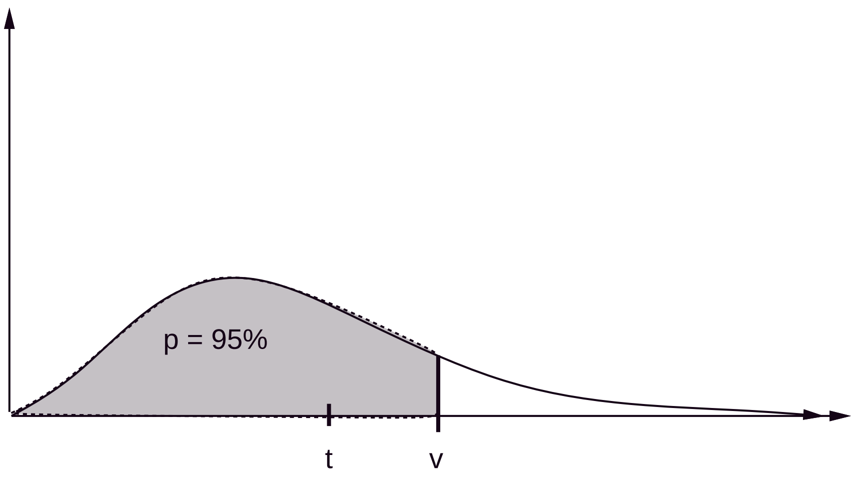
1.5 Test af hypoteser

I dette afsnit skal vi se en metode til statistisk test af hvorvidt to fordelinger har samme middelværdi eller samme varians. I appendix C er yderligere beskrevet hvorledes man tester om to stokastiske variable har samme fordeling. Test af ens middelværdi kan f.eks. udnyttes til afgørelse af om to sæt af observationer kan skelnes eller ej. I de beskrevne metoder benyttes opslag i forskellige fordelingsfunktioner, der ikke er beskrevet i disse noter. Disse kan som regel findes i tabelsamlinger, og er også bestrevet i [9].

Metoden bag statistiske tests er at beregne en såkaldt *teststørrelse* hvis teoretiske fordeling er kendt for en given antagelse. Denne antagelse, hypotesen, betegnes ofte H_0, og kan f.eks. være at middelværdierne for to stokastiske variable er ens. Da antallet af stikprøver er særdeles betydende for sikkerheden på teststørrelsen indgår den (de) tilsvarende antal frihedsgrader ofte i beregning af fordelingen. Det er sjældent muligt at acceptere eller at forkaste nogen hypotese H_0 endegyldigt. Derimod kan vi, givet H_0 og givet et konfidensniveau p bestemme hvorvidt H_0 kan afvises eller ej (på dette konfidensniveau).

Til illustration (og kun til dette formål) kan man alternativ finde konfidensniveauet p ud fra den beregnede teststørrelse t ved opslag i en tabelsamling. Givet at H_0 er sand, vil sandsynligheden for at se en testværdi større end t være lig $1 - p$. Andeledes sagt vil sandsynligheden for fejlagtigt at afvise den korrekte hypotese H_0 være lig $1 - p$.

I praksis fastlægges konfidensniveauet p, dvs. sandsynligheden for fejlagtigt at afvise den korrekte hypotese H_0, før målingen af testværdien. Ved tabelopslag findes den til p svarende grænseværdi v for det aktuelle antal af frihedsgrader. Først nu beregnes teststørrelsen t. Hvis $t > v$ kan vi afvise hypotesen H_0 på konfidensniveauet p. Hvis $t \leq v$ kan hypotesen ikke afvises.



Figur 1.2: Teoretisk fordeling af teststørrelsen under hypotesen H_0 ved f frihedsgrader. Svarende til 95% -konfidensniveauet er grænseværdien v angivet.

Det er klart, at jo større et konfidensniveau man vælger, jo sværere bliver det at afvise hypotesen. Hvis man omvendt vælger p lille vil det blive lettere at afvise H_0. I dette tilfælde er muligheden for fejlagtigt at afvise hypotesen derfor større.

1.5.1 Test af middelværdi

Lad x og y være to stokastiske variable, hvis fordelinger er implicit målt ved stikprøvetagning, således at der er n_x stikprøver af x og n_y stikprøver af y til rådighed. Lad $\hat{m}_x$ hhv. $\hat{m}_y$ være de estimerede middelværdier af x og y. Den numeriske forskel $|\hat{m}_x - \hat{m}_y|$ er ikke noget godt mål for forskellen i middelværdi, fordi denne er afhængig af spredningerne af de to fordelinger. Desuden er en given forskel mellem middelværdierne mere troværdig jo flere stikprøver disse er bestemt ved. Lad σ_D være givet ved:

$$\sigma_D^2 = \frac{\sum_{i=0}^{n_x}(x_i - \hat{m}_x)^2 + \sum_{i=0}^{n_y}(y_i - \hat{m}_y)^2}{n_x + n_y - 2} \left(\frac{1}{n_x} + \frac{1}{n_y}\right) \tag{1.17}$$

Et bedre mål t for hvorvidt to middelværdier virkelig er forskellige fås ved normaliseringen:

$$t = \frac{\hat{m}_x - \hat{m}_y}{\sigma_D} \tag{1.18}$$

Det kan vises at hvis varianserne for x og y ikke er for forskellige, da vil t være fordelt som Students t-fordeling med $n_x + n_y - 2$ frihedsgrader. Students t-fordeling angiver (for det valgte konfidensniveau) grænsen for accept af hypotesen om ens middelværdier. Hvis den målte teststørrelse t er større denne grænse, da kan hypotesen om ens middelværdier afvises (på det givne konfidensniveau).

1.5.2 Test af varians

Lad som i forrige afsnit x og y være stokastiske variable, og lad de estimerede varianser være $Var(x)$ hhv. $Var(y)$. Et mål for hvorvidt fordelingerne for x og y har samme varians er:

$$f = \frac{Var(x)}{Var(y)} \tag{1.19}$$

Hvis f afviger meget fra 1, da indikerer dette at de to varianser faktisk er forskellige. I praksis antages at $Var(x) \geq Var(y)$, således at $f \geq 1$ (ellers ombyttes de to variable). Det kan vises at f er F-fordelt med frihedsgraderne $n_x - 1$ og $n_y - 1$. Hvis den målte teststørrelse f er større end forudsagt af F-fordelingen (på det givne konfidensniveau), da kan antagelsen om ens varianser forkastes (på det givne signifikansniveau).

1.5.3 Andre statistiske tests

I appendix C er der gennemgået en tredie vigtig test, nemlig om to empirisk målte fordelinger (eller en empirisk målt og en teoretisk bestemt) er ens. Her bruges den såkaldte χ^2 -test. Også i denne test indgår antallet af observationer. Hvis der er få, skal der meget til at afvise hypotesen H_0, at de to fordelinger er ens. Hvis der er mange målinger er fordelingerne bedre bestemt, og det bliver lettere at afvise H_0.

1.6 Lineær korrelation

Lineær korrelation er en metode til sammenligning af to ordnede følger af stokastiske variable. Disse kan eksempelvis være to kvantificerede fordelingsfunktioner, eller et sæt af punktpar $(x[i], y[i])$, $i = 1, 2, \ldots, n$, hvor $x[i]$ hhv. $y[i]$ er indicerede koordinater. Teststørrelsen kaldes *korrelationskoefficienten*, og er defineret ved:

$$r = \frac{\sum_i (x[i] - \hat{\mu}_x)(y[i] - \hat{\mu}_y)}{(n-1)\, \hat{\sigma}_x\, \hat{\sigma}_y} \tag{1.20}$$

hvor $\hat{\mu}_x$ hhv. $\hat{\mu}_y$ er de estimerede middelværdier af x hhv. y, og $\hat{\sigma}_x$ hhv. $\hat{\sigma}_y$ er de estimerede spredninger af x hhv. y. Hvis punkterne $(x[i], y[i])$ ligger på en perfekt ret linie med positiv hældning er $r = 1$. Hvis hældningen er negativ er $r = -1$. Hvis punkterne ikke ligger perfekt på en ret linie vil $|r| < 1$. Hvis $r \approx 0$ indikerer dette at de to variable x og y er *ukorrelerede*. Under visse betingelser (normalitet af fordelingerne af x og y mv.) er det muligt at teste en observeret værdi af r mod en fordeling. Det vil føre for vidt her at diskurere dette emne (se evt. [9]).

1.7 Vektorfunktioner, kovarians

I næsten enhver form for statistisk mønstergenkendelse benyttes vektorer af stokastiske variable. Som beskrevet senere knyttes der ofte en featurevektor til hver observation. Vektoren indeholder komponenter, der hver modelleres ved en stokastisk variabel, og som forventes at beskrive et relevant aspekt af observationen. I det nedenstående skal vi kort beskrive de basale termer og definitioner i forbindelse med håndtering af stokastiske vektorer. Senere i noterne skal vi se talrige eksempler på anvendelser.

En stokastisk vektor $\mathbf{x}$ af dimension n er en vektor med n stokastiske variable. Middelværdien af en stokastisk vektor er en vektor hvor hver komponent er middelværdien af den tilsvarende stokastiske variable. *Kovariansen* mellem to stokastiske variable x og y er givet ved:

$$Cov(x, y) = E([x - E(x)][y - E(y)]) \tag{1.21}$$

Varians-kovariansmatricen (ofte blot kaldt kovariansmatricen) for den stokastiske vektor $\mathbf{x}$ er givet ved:

$$\begin{aligned}
\mathbf{C} &= E([\mathbf{x} - \mathbf{m_x}][\mathbf{x} - \mathbf{m_x}]^t) & (1.22)\\
&= E\left(\begin{bmatrix} x_1 - m_1 \\ x_2 - m_2 \\ \vdots \\ x_n - m_n \end{bmatrix} [(x_1 - m_x), (x_2 - m_2), \ldots, (x_n - m_x)]\right) \\
&= \begin{bmatrix} E([x_1 - m_1][x_1 - m_1]) & E([x_1 - m_1][x_2 - m_2]) & \cdots & E([x_1 - m_1][x_n - m_n]) \\ E([x_2 - m_2][x_1 - m_1]) & E([x_2 - m_2][x_2 - m_2]) & \cdots & E([x_2 - m_2][x_n - m_n]) \\ \vdots & & \ddots & \vdots \\ E([x_n - m_n][x_1 - m_1]) & E([x_n - m_n][x_2 - m_2]) & \cdots & E([x_n - m_n][x_n - m_n]) \end{bmatrix} \\
&= \begin{bmatrix} VAR(x_1) & COV(x_1, x_2) & \cdots & COV(x_1, x_n) \\ COV(x_2, x_1) & VAR(x_2) & \cdots & COV(x_2, x_n) \\ \vdots & & \ddots & \vdots \\ COV(x_n, x_1) & COV(x_n, x_2) & \cdots & VAR(x_n) \end{bmatrix} & (1.23)
\end{aligned}$$

hvor m_i er middelværdien af den i'te komponent af vektoren $\mathbf{x}$. Bemærk at diagonalelementerne af kovariansmatricen er lig varianserne af de enkelte stokastiske variable. De øvrige elementer kaldes kovarianser. Kovariansmatricen har dimmension $n \times n$, og ses at være symmetrisk. Vi skal udelukkende betragte reelle stokastiske vektorvariable, og kan derfor udnytte de mange egenskaber som gælder for reelle symmetriske matricer, eksempelvis at matricen er diagonaliserbar (ved en ortonormal matrice), og at egenværdierne er reelle.

Kovariansmatricen kan også skrives:

$$\mathbf{C} \;=\; E(\mathbf{x}\mathbf{x}^t) \;-\; E(\mathbf{x})\mathbf{m_x}^t \;-\; \mathbf{m_x}E(\mathbf{x}^t) \;+\; \mathbf{m_x}\mathbf{m_x}^t \;=\; \mathbf{S} \;-\; \mathbf{m_x}\mathbf{m_x}^t \tag{1.24}$$

hvor

$$\mathbf{S} \;=\; E(\mathbf{x}\mathbf{x}^t) \;=\; \begin{bmatrix} E(\mathbf{x}_1\mathbf{x}_1) & \cdots & E(\mathbf{x}_1\mathbf{x}_n) \\ \vdots & & \vdots \\ E(\mathbf{x}_1\mathbf{x}_n) & \cdots & E(\mathbf{x}_n\mathbf{x}_n) \end{bmatrix} \tag{1.25}$$

Matricen $\mathbf{S}$ kaldes *autokorrelationsmatricen* .

I visse tilfælde dekomponeres $\mathbf{C} = \mathbf{\Lambda R \Lambda}$, hvor:

$$\mathbf{\Lambda} \;=\; \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & & \sigma_n \end{bmatrix} \tag{1.26}$$

og

$$\mathbf{R} \;=\; \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{1n} & 1 & & \vdots \\ \vdots & & \ddots & \vdots \\ \rho_{1n} & \cdots & & 1 \end{bmatrix} \tag{1.27}$$

hvor $|\rho_{ij}| \leq 1$, og hvor $\rho_{ij} = \rho_{ji}$. Elementerne c_{ij} af kovariansmatricen $\mathbf{C}$ kan således skrives:

$$c_{ij} \;=\; \begin{cases} \sigma_i^2 & \text{hvis } i = j \\ \rho_{ij}\sigma_i\sigma_j & \text{hvis } i \neq j \end{cases} \tag{1.28}$$

Størrelsen σ_i kaldes spredningen (eller standard afvigelsen) af x_i, og ρ_{ij} kaldes *korrelationskoefficienten* mellem de stokastiske variable x_i og x_j. Matricen $\mathbf{R}$ kaldes *korrelationsmatricen*, og generaliserer den sædvanlige korrelationskoefficient fra ligning (1.20). Matricen $\mathbf{R}$ indeholder den essentielle information om hvorledes de stokastiske variable er indbyrdes relateret.

Ved analyse af et sæt vektordata er konstruktion af kovariansmatricen ofte noget af det første der foretages. Som vi skal se senere giver en analyse af denne matrice information om graden af lineær relation mellem komponenterne i vektoren (hvis ρ_{ij} er stor for $i \neq j$). Eksempelvis vil det være muligt at undersøge om dimensionaliteten af vektoren kan reduceres uden væsenligt informationstab (ved bortkastning af en eller flere af vektorkomponenterne). Iøvrigt beskriver matricen hvor stor variation hver af komponenterne udviser. Hvis σ_i er lille, da beskriver den i'te komponent et fællestræk for observationerne. Hvis σ_i er stor, er det måske muligt at differentiere mellem observationerne på basis af denne information.

Kovariansen kan estimeres direkte ud fra (1.24) ved erstatning af den forventede værdi med et gennemsnit over observationerne. Dette estimat kan (som for variansestimatet) vises at være biased. Et unbiased estimat er:

$$\hat{\mathbf{C}} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}-\hat{\mathbf{m}})(\mathbf{x}-\hat{\mathbf{m}})^t \tag{1.29}$$

Eksempel Lad (r, g, b) være mængden af rødt, grønt og blåt lys, som et kamera har registreret i et billede af dimension 4×4. De 16 registrerede værdier var:

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
r	1	1	2	3	1	2	3	3	2	3	4	4	4	4	5	6
g	7	8	9	10	8	9	10	11	10	10	13	14	12	12	15	16
b	4	4	4	8	4	4	8	8	2	4	8	8	1	2	4	8

Middelværdivektoren bliver $(3.0000, 10.8750, 5.0625)$. Kovariansmatricen kan udregnes til:

$$\begin{aligned}
\hat{\mathbf{C}} &= \begin{bmatrix} 3.200 & 3.667 & 1.666 \\ 3.667 & 6.783 & 2.075 \\ 1.666 & 2.075 & 6.329 \end{bmatrix} \qquad (1.30) \\
&= \begin{bmatrix} 1.789 & 0 & 0 \\ 0 & 2.604 & 0 \\ 0 & 0 & 2.516 \end{bmatrix} \begin{bmatrix} 1.000 & 0.787 & 0.370 \\ 0.787 & 1.000 & 0.317 \\ 0.370 & 0.317 & 1.000 \end{bmatrix} \begin{bmatrix} 1.789 & 0 & 0 \\ 0 & 2.604 & 0 \\ 0 & 0 & 2.516 \end{bmatrix}
\end{aligned}$$

Det ses at varianserne er af samme størrelsorden, den røde lidt mindre end de to andre. Korrelationskoefficienten mellem de røde og grønne farver er forhodsvis stor, hvorimod korrelationenskoefficienterne mellem den blå og den grønne hhv. den røde farve er forholdsvis lille.

Eksempel slut

Stokastiske vektorfunktioner er naturligvis, som enkelte stokastiske variable, beskrevet ved en sandsynlighedsfordeling. Hvis de enkelte variable i vektoren er uafhængige fremkommer denne, som sædvanligt, ved multiplikation af fordelingerne for de enkelte variable. Denne situation er imidlertid atypisk. Vi skal her kun beskrive én, meget benyttet, multivariat fordeling, nemlig normalfordelingen. Lad $\mathbf{x}$ være en stokastiske vektor af dimension n, lad middelværdien af $\mathbf{x}$ være $\mathbf{m}$, og lad kovariansmatricen være givet ved $\mathbf{C}$. Da er normalfordelingen givet ved:

$$\begin{aligned}
p(\mathbf{x}) &= \frac{1}{(2\pi)^{n/2}|\mathbf{C}|^{1/2}} \quad exp[-\frac{1}{2}(\mathbf{x}-\mathbf{m})\mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})^t] \qquad (1.31) \\
&= \frac{1}{(2\pi)^{n/2}|\mathbf{C}|^{1/2}} \quad exp[-\frac{1}{2}d^2(\mathbf{x})] \qquad (1.32)
\end{aligned}$$

Relationen til den endimensionale normalfordeling ses let i det tilfælde hvor alle kovarianser er lig 0 (dvs. at $\mathbf{C}$ er en diagonalmatrice). I dette tilfælde reducerer (1.31) til:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\sigma_1\sigma_2\cdots\ \sigma_n} \quad exp\left[-\frac{1}{2}\sum_{i=1}^{n}\frac{(\mathbf{x}-\mathbf{m})^2}{\sigma_i^2}\right] \tag{1.33}$$

Størrelsen $d^2(\mathbf{x})$ i udtrykket for den flerdimensionale normalfordeling angiver et normeret udtryk for afstanden mellem observationen $\mathbf{x}$ og middelværdien $\mathbf{m}$, her betragtet som punkter i en n-dimensionalt rum. Udtrykket kaldes også for *Mahalanobis afstanden.* Vi skal senere i noterne udnytte den flerdimensionale normalfordeling, og Mahalanobis afstanden, intensivt.

Som for to fordelinger, hver af en enkelt stokastisk variabel, findes der metoder til sammenligning af flerdimensionale fordelinger. To populære afstandsmål er *Bhattacharyya* afstanden B og *divergensen* D. Lad to normalfordelinger f_1 og f_2 være karakteriseret at deres middelværdier $\mathbf{m}_1$ og $\mathbf{m}_2$, samt deres kovariansmatricer $\mathbf{C}_1$ og $\mathbf{C}_2$. Da er:

$$B(f_1, f_2) \;=\; \frac{1}{8}(\mathbf{m}_2 - \mathbf{m}_1)^t \left(\frac{\mathbf{C}_1 + \mathbf{C}_2}{2}\right)^{-1} (\mathbf{m}_2 - \mathbf{m}_1) \;+\; \frac{1}{2} ln \frac{\frac{1}{2}|\mathbf{C}_1 + \mathbf{C}_2|}{\sqrt{|\mathbf{C}_1|} + \sqrt{|\mathbf{C}_2|}} \tag{1.34}$$

og

$$D(f_1, f_2) \;=\; Tr\left[(\mathbf{C}_1 - \mathbf{C}_2)(\mathbf{C}_2^{-1} - \mathbf{C}_1^{-1})\right] \;+\; \frac{1}{2} Tr\left[(\mathbf{C}_2^{-1} + \mathbf{C}_1^{-1})(\mathbf{m}_1 - \mathbf{m}_2)(\mathbf{m}_1 - \mathbf{m}_2)^t\right] \tag{1.35}$$

Det kan bemærkes at hvis $\mathbf{C}_1 = \mathbf{C}_2$, så er $d^2 = D$. I dette tilfælde er disse to afstande altså ækvivalente. Vi skal ikke her diskutere disse mål yderligere. Den interesserede læser henvises til [5], der mere fyldigt analyserer de nævnte mål.

Man har ofte behov for at transformere en stokastisk vektor $\mathbf{x}$ til en ny stokastisk vektor $\mathbf{y}$. Givet fordelingen f_x af $\mathbf{x}$ er det derfor interessant at kunne beregne fordelingen f_y af $\mathbf{y}$. Lad dimensionen af $\mathbf{x}$ være n. Eksempler på transformationer er da: $\sum_{i=1}^n x_i$, $\sum_{i=1}^n x_i^2$, $\max_i\{x_i\}$, etc. Teknisk er det en nøvendighed at antage, at dimensionen af $\mathbf{y}$ er lig dimensionen af $\mathbf{x}$. Lad derfor:

$$\begin{aligned} y_1 &= \phi_1(x_1, x_2, \ldots, x_n) \\ y_2 &= \phi_2(x_1, x_2, \ldots, x_n) \\ &\vdots \\ y_n &= \phi_n(x_1, x_2, \ldots, x_n) \end{aligned} \tag{1.36}$$

Vi kan skrive ovenstående som $\mathbf{y} = \phi(\mathbf{x})$. Hvis ϕ er bijektiv, med den omvendte afbilding givet ved $\mathbf{x} = \psi(\mathbf{y})$, da er f_y bestemt ved:

$$f_y(\mathbf{y}) \;=\; f_x(\psi(\mathbf{y}))\left|\frac{\partial(\mathbf{x})}{\partial(\mathbf{y})}\right| \tag{1.37}$$

for alle $\mathbf{y}$ i værdimængden for ϕ. Bemærk at der tages numerisk værdi af den anden faktor i (1.37). Størrelsen

$$\frac{\partial(\mathbf{x})}{\partial(\mathbf{y})} \;=\; det\left\{\frac{\partial x_i}{\partial y_j}\right\}_{i,j=1,..,n} \;=\; \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \cdots & \frac{\partial x_2}{\partial y_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_n} \end{vmatrix} \tag{1.38}$$

kaldes *Jacobis funktionaldeterminant* eller blot *Jacobi-determinanten.*

Eksempel
Lad x_1, og x_2 være to stokastiske uafhængige normalfordelte variable begge med middelværdi 0 og spredning 1. Vi ønsker nu at bestemme fordelingen af $y_1 = x_1/x_2$, og tilføjer derfor hjælpevariablen $y_2 = x_2$. Den omvendte afbildning er bestemt ved:

$$\begin{aligned} x_1 &= y_1 y_2 \\ x_2 &= y_2 \end{aligned}$$

Jacobi-determinenten bliver:

$$\frac{\partial(\mathbf{x})}{\partial(\mathbf{y})} = \begin{vmatrix} y_2 & y_1 \\ 0 & 1 \end{vmatrix} = y_2$$

Da x_1 og x_2 er uafhængige fås fordelingsfunktionen for $\mathbf{x}$ til:

$$f(\mathbf{x}) = \frac{1}{2\pi} exp(-\frac{1}{2}[x_1^2 + x_2^2])$$

Ved indsættelse af de udledte størrelser fås:

$$g(\mathbf{y}) = \frac{1}{2\pi} exp(-\frac{1}{2}[(y_1 y_2)^2 + y_2^2]) \; |y_2|$$

Da domænet for alle de involverede stokastiske variable er hele den reelle akse får vi fordelingen af y_1 ved at integrere afhængigheden af y_2 ud, dvs:

$$\begin{aligned} g(y_1) &= 2 \int_0^\infty \frac{1}{2\pi} y_2 \; exp(-\frac{1}{2}[(y_1 y_2)^2 + y_2^2]) dy_2 \\ &= \frac{1}{\pi} \frac{1}{1 \; + \; y_1^2} \end{aligned}$$

Eksempel slut

1.8 Opgaver

1. Vis at konstanten 1.4826 i ligning (1.15) fremkommer som beskrevet i teksten. Brug at:

$$\Theta(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt$$

   Vink: Find medianen af $|v|$, hvor v er en stokastisk normalfordelt variabel med middelværdi 0.

2. Betragt følgende 15 tal:

   4, 3, 7, 1, 2, 3, 6, 5, 5, 4, 2, 9, 1, 3, 2

Beregn estimater af den forventede værdi ved brug af den sædvanlige definition af middelværdi ved brug af medianmetoden, og ved brug af den 0.2-trimmede middelværdi. Beregn estimater af spredningen ved den sædvanlige definition samt ved MeanAD-metoden. Erstat nu tallet 6 med outlieren 12, og gentag udregningen. Erstat yderligere tallet 3 med outlieren 0, og gentag udregningen.

3. Betragt nedenstående sekvens af målinger:

4	6	8	6	7	3	9	1	0	11
6	9	7	2	8	8	12	3	7	5
6	6	7	9	2	9	2	12	3	7
22	4	15	0	6	5	7	2	2	5
1	4	5	4	8	10	13	10	11	10

Beregn middelværdien og spredningen af målingerne.

4. Betragt nedenstående sekvens af målinger:

3	0	8	4	3	2	2	0	1	4
6	9	0	1	5	1	10	0	0	1
0	1	0	4	5	7	16	0	9	8
6	2	3	1	0	1	10	7	6	4
11	5	2	3	2	1	12	0	0	3

Beregn middelværdien og spredningen af målingerne.

5. Test hvorvidt målingerne vist i opgave 3 og opgave 4 har samme a) middelværdi, b) varians.

6. Betragt nedenstående sekvens af målinger:

1	2	3	4	5	6	2	7	7	6
2	3	4	5	5	7	10	8	7	5
4	5	6	7	5	8	12	1	6	9
8	9	0	8	9	10	6	5	4	17
6	7	5	6	9	2	20	5	4	9
3	6	10	11						

Afgør om ovenstående målinger samt målingerne vist i opgave 3 har a) samme middelværdi, b) samme varians,

7. Betragt nedenstående to indicerede datasæt x og y.

i	1	2	3	4	5	6	7	8	9	10	11	12
x	3	0	8	4	3	2	2	0	1	4	5	12
y	4	-4	17	8	3	2	1	-5	0	7	9	3

Bestem den sædvanlige korrelationskoefficient mellem sættet bestående af x-værdierne og sættet af y-værdier, samt Spearmans rank-order korrelationskoefficient. Er der grund til at tro at de to datasæt er korrelerede.

8. Lad x og y være to uafhængige normalfordelte stokastiske variable med middelværdi 0 og med ens spredning σ. Hvad er fordelingen af $x^2 + y^2$.

9. Lad x og y være to uafhængige normalfordelte stokastiske variable med middelværdi 0 og med ens spredning σ. Hvad er fordelingen af $x + y$, og $x - y$.

10. Lad x være en eksponentialfordelt stokastisk variabel med parameter 0.5 defineret på intervallet [0:100]. Bestem den transformation ϕ, der transformerer x til en ligefordelt stokastisk variabel y.

11. Betragt målingerne fra opgave 6, og antag at disse kan modelleres som stikprøver af en heltallig stokastisk variabel x, med værdiområde [0:20]. Bestem den transformation ϕ, der transformerer x til en "ligefordelt" heltallig stokastisk variabel y. Foretag transformationen på målingerne.

Kapitel 2

Selektion af features

Det er et grundlæggende problem i dataanalyse at de målinger som et givet apparat leverer ikke nødvendigvis er de mest hensigtsmæssige, i rå form, til karakterisering af den proces, som det er hensigten at forstå eller modellere.

Til eksempel består et multispektralt satellitbillede af et antal samtidigt optagne billeder. Hvert billede er dannet af en sensor, der er følsom inden for et område af det elektromagnetiske spektrum. Til ethvert billedelement svarer således en vektor $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ af målinger, hvor n er antallet af sensorer. Da de spektrale bånd overlapper hinanden en del, og da visse materialer reflekterer lys over et bredt spektrum af frekvenser, er informationsindholdet i billederne meget redundant (elementerne i observationsvektoren $\mathbf{x}$ er korrelerede). Anderledes sagt vil informationen af en observation $\mathbf{x}$ være indeholdt i et underrum af $\mathcal{R}^n$. Før videre analyse foretages kan det derfor være hensigtsmæssigt at reducere dimensionaliteten af observationerne, således at det meste af informationen bevares (og således at en stor del af støjen elimineres).

I visse situatuationer beskriver de aktuelle observationer langt mere end det er formålet med analysen at beskrive/modellere. I disse tilfælde er det ofte hensigtsmæssigt indledende at ekstrahere den brugbare information. Dette kan gøres ved, til at der til hver observation (eller grupper af observationer) knyttes en *feature-vektor*, således at hver feature beskriver en relevant egenskab eller karakteristika. I visse tilfælde er det forholdsvist let at udvælge de relevante features, i andre tilfælde kan dette være et hovedproblem. Det kan være særdeles vanskeligt af definere features, der udelukkende beskrive relevante aspekter af de observerede data, og ignorerer den irrelevante information.

En prototype på et system til genkendelse og klassifikation består af 4 faser. I første fase registreres og grupperes måledata i k vektorer. I eksemplet med de multispektrale billeder består denne fase i, for hver billedelement, at konstruere observationsvektoren $\mathbf{x}$. I anden fase ekstraheres features ved til hver af de k vektorer $\mathbf{x} \in \mathcal{R}^n$, at bestemme en featurevektor $\mathbf{y} \in \mathcal{R}^p$, hvor p er antallet af features. I tredie fase foretages en klyngeanalyse på alle (eller en delmængde af) featurevektorene. I klyngeanalysen identificeres et sæt af m prototyper (eller klasser). Hver klasse antages at beskrive et objekt. Hver klasse tilknyttes en (ofte statistisk) beskrivelse af position og udstrækning af klassen i featurerummet. Til slut tilknyttes hver af de k featurevektorer et objektnummer. Dette sker ved at sammenligne hver af de k featurevektorer med beskrivelserne af de m klasser, og ved at udvælge den klasse der, i en nærmere angivet forstand, ligner bedst. Typisk foretages klyngeanalysen på et sæt af træningsdata. Klassifikationen kan anvendes på data der er opsamlet på et senere tidspunkt.

Nedenstående introduceres til nogle klassiske metoder til ekstraktion af features. Metoder til klyngeanalyse beskrives i kapitel 4. Metoder til klassifikation beskrives i kapitel 5.

2.1 Principalkomponentanalyse

Principalkomponent analyse er en klassisk metode til analyse af dimensionaliteten af et sæt observationer. Principalkomponent analyse er også kendt som Karhuen-Loeve transformationen, og som Hotelling transformationen. Den grundlæggende idé bag metoden er en egenværdianalyse af kovariansmatricen for observationerne. Metoden er således baseret på de statistiske egenskaber ved observationsvektoren $\mathbf{x}$. Hver enkeltmåling x_i betragtes som en stokastisk variabel.

Lad $\mathbf{x}$ være en observationsvektor med n elementer. Kovariansmatricen $\mathbf{C_x}$ for $\mathbf{x}$ har størrelse $n \times n$ og defineres ved:

$$\mathbf{C_x} \;=\; E\{(\mathbf{x}-\mathbf{m_x})(\mathbf{x}-\mathbf{m_x})^t\} \;=\; E\{\mathbf{x}\mathbf{x}^t\} - \mathbf{m_x}\mathbf{m_x}^t \tag{2.1}$$

$E\{\cdot\}$ angiver statistisk forventet værdi, og:

$$\mathbf{m_x} \;=\; E\{\mathbf{x}\} \tag{2.2}$$

Vektoren $\mathbf{m_x}$ har således længden n. Lad nu $\mathbf{e}_i$ hhv. λ_i være den i'te egenvektor hhv. egenværdi for $\mathbf{C_x}$ (i = 1,2,...,n), og lad $\mathbf{A}$ være en transformationsmatrice hvis rækker er bestemt ved egenvektorerne $\mathbf{e}_i$.

$$\mathbf{A} \;=\; \begin{pmatrix} e_{11} & e_{12} & \cdots & e_{1n} \\ e_{21} & e_{22} & \cdots & e_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ e_{n1} & e_{n2} & \cdots & e_{nn} \end{pmatrix} \;=\; \begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \\ \vdots \\ \mathbf{e}_n \end{pmatrix} \tag{2.3}$$

således at e_{ij} er den j'te komponent af den i'te egenvektor. Karhuen-Loeve transformationen er da givet ved:

$$\mathbf{y} \;=\; \mathbf{A}(\mathbf{x}-\mathbf{m_x}) \tag{2.4}$$

Denne transformation har to interessante egenskaber, nemlig:

$$\mathbf{m_y} \;=\; E\{\mathbf{y}\} \;=\; E\{\mathbf{A}(\mathbf{x}-\mathbf{m_x})\} \;=\; \mathbf{A}E\{\mathbf{x}\} - \mathbf{A}\mathbf{m_x} \;=\; 0 \tag{2.5}$$

og

$$
\begin{aligned}
\mathbf{C_y} &= E\{(\mathbf{y}-\mathbf{m_y})(\mathbf{y}-\mathbf{m_y})^t\} \\
&= E\{\mathbf{y}\mathbf{y}^t\} \\
&= E\{(\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{m_x})(\mathbf{A}\mathbf{x}-\mathbf{A}\mathbf{m_x})^t\} \\
&= E\{\mathbf{A}(\mathbf{x}-\mathbf{m_x})(\mathbf{x}-\mathbf{m_x})^t\mathbf{A}^t\} \\
&= \mathbf{A}\mathbf{C_x}\mathbf{A}^t \\
&= diag(\lambda_1, \lambda_2, \ldots, \lambda_n)
\end{aligned} \tag{2.6}
$$

Middelværdien $\mathbf{m_y}$ af transformationskoefficienterne er altså lig nulvektoren. Fra lineær algebra genkender vi udtrykket for $\mathbf{C_y}$ som diagonaliseringen af $\mathbf{C_x}$. Dette betyder at $\mathbf{C_y}$ har formen:

$$
\mathbf{C_y} = \begin{pmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_n \end{pmatrix} \tag{2.7}
$$

At kovariansmatricen for koordinaterne $\mathbf{y}$ til transformationen er en diagonalmatrice betyder, at koordinaterne er fuldstædig dekorrelerede. Størrelsen af egenværdien λ_i angiver hvor stor en del af energien i de oprindelige observationer, der er knyttet til underrummet svarende til den i'te egenvektor. Sædvanligvis ordnes egenvektorerne således at $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Det kan yderligere bemærkes at λ_i er lig variansen af det i'te elemenet af $\mathbf{y}$. Da $\mathbf{C_x}$ er reel og symmetrisk vil $\mathbf{A}$ altid være en ortonormal matrice. Heraf følger at $\mathbf{A}^{-1} = \mathbf{A}^t$. Den inverse transformation er derfor givet ved:

$$
\mathbf{x} = \mathbf{A}^t\mathbf{y} + \mathbf{m_x} \tag{2.8}
$$

Antag at vi ønsker en rekonstruktion $\hat{\mathbf{x}}$ på basis af de k første egenvektorer af $\mathbf{C_x}$. Vi har da.

$$
\hat{\mathbf{x}} = \mathbf{A}_k^t\mathbf{y} + \mathbf{m_x} \tag{2.9}
$$

Ved udregning af kvadratfejlen er det praktisk at udtrykke denne som sporet (Tr) af det ydre produkt af differensen $(\mathbf{x} - \hat{\mathbf{x}})$ i stedet for, som normalt, det indre produkt. Vi kan opskrive kvadratfejlen som:

$$
\begin{aligned}
e_k^2 &= ||\mathbf{x} - \hat{\mathbf{x}}||^2 \\
&= ||(\mathbf{A}^t - \mathbf{A}_k^t)\mathbf{y}||^2 \\
&= Tr[E\{\left[(\mathbf{A}^t - \mathbf{A}_k^t)\mathbf{y}\right]\left[(\mathbf{A}^t - \mathbf{A}_k^t)\mathbf{y}\right]^t\}] \\
&= Tr[E\{(\mathbf{A}^t - \mathbf{A}_k^t)\mathbf{y}\mathbf{y}^t(\mathbf{A} - \mathbf{A}_k)\}] \\
&= Tr[(\mathbf{A}^t - \mathbf{A_k}^t)\mathbf{A}\mathbf{C}_x\mathbf{A}^t(\mathbf{A} - \mathbf{A}_k)] \\
&= Tr[(\mathbf{A} - \mathbf{A}_k)^t\mathbf{C}_y(\mathbf{A} - \mathbf{A}_k)] \\
&= \sum_{j=k+1}^{n} \lambda_j
\end{aligned} \tag{2.10}
$$

Da egenværdierne λ_i antages sorteret efter aftagende værdi, ses at transformationen optimalt pakker mest mulig energi i koordinaterne svarende til de første egenvektorer.

I praksis benyttes metoden til dels at reducere dimensionaliteten af observationsvektoren, dels til dekorrelation af de resterende koordinater. Som featurevektor benyttes derfor $\mathbf{y} \in \mathcal{R}^p$, hvor $p \leq n$. Dimensionen reduceres ved at ignorere de transformerede koordinater (rækker af matricen $\mathbf{A}^t$) hørende til de mindste egenværdier. Antallet af ignorerede dimensioner kan bestemmes på forhånd, ud fra et krav om bevarelse af en vis procentdel af energien i observationerne, eller ud fra en manuel inspektion af de transformerede koordinatværdier.

2.2 Andre transformationer

Principalkomponent transformationen er interessant fordi den sikrer nogle statistiske egenskaber ved featurevektoren. Der er imidlertid ofte andre egenskaber, som kan relevante. Som eksempel skal vi her tænke på observationer, der er fremkommet ved sampling af et kontinuert signal over tid. Observationsvektoren $\mathbf{x}[i]$ er her endimensional, og index i angiver samplingtidspunktet. Eksempler på data af den beskrevne form ses ofte ved analyse af lyd, af allehånde former for elektriske signaler, f.eks. medicinsk overvågningsudstyr vibrationsmålere etc.

I en *frekvensanalyse* af et tidssamplet signal foretages ofte en *fouriertransformation*. Fouriertransformationen beskriver, som principalkomponent transformationen, et baseskift, og er lineær. For den diskrete fouriertransformationen er basefunktionerne givet ved de komplekse eksponentialfunktioner $exp(-2\pi i x u/N)$, hvor u er frekvensvariablen, $i = \sqrt{-1}$, og N er antal stikprøver. Frekvensanalyse og herunder fouriertransformationen beskrives i detalje i an andel del af kurset. Fouriertransformationen kan være nyttig, hvis den relevante information er knyttet til en repetition af mønster i observationsvektoren. Store værdier i det transformerede signal indikerer en repetition af et mønster, og fortæller hvor hyppigt mønstret gentages. Fouriertransformation er især relevant hvis repetitionshyppigheden rummer den interessante information. Dette er f.eks. tilfældet ved overvågning af hjerterymen på en patient under operation. Transformationen er også klassisk ved analyse af tale og musik, herunder metoder til talegenkendelse.

Principalkomponent transformationen og fouriertransformationen er naturligvis ikke de eneste interessante linære transformationer. Der findes en lang række af lineære transformationer, alle beskrevet ved ligningerne:

$$T[u] = \sum_{x=0}^{N-1} f[x]g[x,u] \tag{2.11}$$

$$f[x] = \sum_{u=0}^{N-1} T[x]h[x,u] \tag{2.12}$$

hvor N er dimensionen af såvel observationsvektoren $f[x]$ som transformations *koefficienterne* $T[u]$. Funktionen g kaldes for *kernen* for transformationen. Det ses at hver koefficient $T[u]$ beregnes ved projektion af f på en basisfunktion $g_u[x] = g[x, u]$. Valget af gode features svarer således til valget af gode basisfunktioner. Det er vanskeligt at give generelle råd for valg af disse, dog ønskes sædvangigvis at de er ortogonale.

2.2.1 Ikke-lineære transformationer

Principalkomponent transformantionen og fouriertransformationen er lineære transformationer. I visse anvendelser er ikke-lineære transformationer bedre i stand til at isolere den relevante information. Nedenstående gives et sådant eksempel.

Ved kassifikation of objekter ud fra et farvebillede er det ofte hensigtsmæssigt at isolere farvetonen. Sædvanligvis består observationsvektoren (for hver billedelement) af en rød, en grøn, og en blå komponent (RGB-format), dvs. $\mathbf{x} = (R, G, B)$. Værdierne af de tre komponenter er ofte kraftigt korreleret, og beskriver kun inddirekte hvilken farve der er tale om. Intensiteten I defineres som længden af vektoren $\mathbf{x}$. En alternativ repræsentation kan derfor være beskrevet ved den ulineære transformation:

$$\begin{aligned} I &= R + G + B \\ r &= R/I \\ g &= G/I \end{aligned}$$

Farveinformationen er her reduceret til koordinaterne r og g. Denne transformation er imidlertid sjældent hensigtsmæssig. I stedet benyttes en anden ulineær transformation, hvor koordinaterne (ud over intensiteten I) direkte angiver farvetoningen H og farvemætningsgraden S. Vi skal ikke her gennemgå udledningen, men blot notere at H kan beregnes ved:

$$H = \cos^{-1}\left\{ \frac{\frac{1}{2}\left[(R-G) + (R-B)\right]}{\sqrt{(R-G)^2 + (R-B)(G-B)}} \right\} \tag{2.13}$$

Ovenstående (meget) ulineære transformation isolerer således den relevante information af observationsvektoren $\mathbf{x}$, og letter enhver efterfølgende analyse hvor klassifikation på basis af farvetoning alene skal foretages.

2.3 Kvalitetsvurdering af features

Et væsentlig egenskab ved en feature er, at den beskriver en relevant egenskab ved den klasse af objekter, som man ønsker at identificere, og at den er ufølsom over for støj og irrelevante egenskaber. Disse krav er svære at opfylde samtidigt. Ofte vil én eller få features ikke være tilstrækkeligt til at separere de forskellige objekter, som man ønsker at differentiere imellem, hvilket kan nødvendiggøre brugen af mange features. I dette tilfælde er det hensigtsmæssigt at vælge sine features således, at de beskriver forskellige aspekter af objekterne. Med andre ord bør de enkelte features vælges således at korrelationen imellem disse er lille. Til dette formål kan en principalkomponentanalyse være brugbar.

Ved konstruktion af en klassifikationsprocedure er det ofte nødvendigt at afprøve et stort antal features. Til dette formål benyttes et sæt af observationsvektorer $\mathbf{x}$, hvorom deres objekt tilhørsforhold er kendt (f.eks. ved manuel klassifikation). Der formuleres derefter et sæt af (ikke nødvendigvis lineære) transformationer $f_i : \mathbf{x} \rightarrow \mathbf{y}$, hvorved et sæt af featurevektorer $\mathbf{y}$ bestemmes. Opgaven er nu at måle kvaliteten af en delmængde af de mulige features. Først grupperes vektorerne $\mathbf{y}$ i k grupper efter den klasse de (ved korrekt klassifikation) er repræsentanter for. For hver klasse k ønskes variationen af featureværdierne så lille som muligt. Desuden ønskes afstanden mellem featureværdierne for ethvert par af klasser så stor som muligt.

Lad $\mathbf{m}_k$ være middelværdivektoren og lad $\mathbf{C}_k$ være kovariansmatricen for for den k'te klasse. Sættet af features er godt hvis alle kovaransmatricer er tæt på at være diagonalmatricer, og hvis varians-kovarians-værdierne er forholdsvist små (se senere). Klasserne er godt sepereret, hvis der om ethvert par at klasser $i \neq j$ gælder at $|\mathbf{m}_i - \mathbf{m}_j|$ er forholdsvis stor. Hvis featureværdierne for hver af de k populationer antages at være normelfordelt kan ovenstående formaliseres på en kompakt form. Antagelsen om normalitet er, for hver klasse k, at:

$$p_k(\mathbf{x}) \;=\; \frac{1}{(2\pi)^{n/2}|\mathbf{C}_k|^{1/2}}\, exp\left[-\frac{1}{2}(\mathbf{x}-\mathbf{m}_k)^t\mathbf{C}_k^{-1}(\mathbf{x}-\mathbf{m}_k)\right] \tag{2.14}$$

hvor n er dimensionen af featurevektorene. I dette tilfælde er *divergensen* af to populationer i og j defineret ved:

$$Q_{ij} \;=\; \frac{1}{2}Tr\left[(\mathbf{C}_i-\mathbf{C}_j)(\mathbf{C}_j^{-1}-\mathbf{C}_i^{-1})\right] \;+\; \frac{1}{2}Tr\left[(\mathbf{C}_j^{-1}+\mathbf{C}_i^{-1})(\mathbf{m}_i-\mathbf{m}_j)(\mathbf{m}_i-\mathbf{m}_j)^t\right] \tag{2.15}$$

Det vil føre for vidt her, at udlede dette udtryk. Divergensen Q_{ij} kan betragtes som en generaliseret afstand mellem to populationer. Vi ønsker at vælge vores sæt af features således at denne afstand er stor for ethvert par af klasser. Et samlet kvalittsmål for sættet af features er derfor:

$$Q \;=\; \sum_{i=1}^{k}\sum_{j=i+1}^{k} Q_{ij} \tag{2.16}$$

Udtrykket for divergensen er en anelse kompliceret. For at øge forståelsen kan vi iagttage nogle specialtilfælde. Antag først at kovariansmatricerne for de to populationer er ens, dvs. $\mathbf{C}_i = \mathbf{C}_j = \mathbf{C}$. Da reduceres udtrykket for divergensen til den såkaldte *Mahalanobis afstand.*

$$Q_{ij} \;=\; (\mathbf{m}_i-\mathbf{m}_j)^t\mathbf{C}^{-1}(\mathbf{m}_i-\mathbf{m}_j) \tag{2.17}$$

Bemærk at Mahalanobis afstanden her angiver et mål for afstanden mellem to fordelinger og ikke (som sædvanligt) afstanden fra et punkt til en fordeling.

Hvis sættet af features yderligere er dekorrelerede vil $\mathbf{C} = diag(c_1, c_2, \dots c_n)$ være en diagonalmatrice, og:

$$Q_{ij} \;=\; \sum_{l=1}^{n}\frac{(\mathbf{m}_i(l)-\mathbf{m}_j(l))^2}{c_l} \tag{2.18}$$

I dette tilfælde ses at Q_{ij} maksimeres når en vægtet sum af kvadratafstandene mellem centrene $\mathbf{m}_k$ maksimeres. Vægtene er her omvendt proportionale med variansen af den enkelte feature. Antages, som det simpleste tilfælde, at alle varianser er ens, dvs. at $\mathbf{C}_i = \mathbf{C}_j = \mathbf{I}$ fås at Q bliver summen af de euclidiske afstande mellem samtlige par af klasser (klyngecentre).

Vi har hermed udledt et udtryk, der under de nævnte antagelser (primært om normalitet), angiver et kvalitetsmål for et sæt af features. Optimering af dette udtryk over samtlige tænkelige sæt af features er naturligvis ingen enkel sag. Derimod angiver udtrykket de komponenter (kovariansmatricerne og middelværdivektorene), der bør undersøges nærmere ved udvælgelsen af et godt sæt af features.

2.4 Opgaver

1. Betragt følgende 10 målinger af tre stokastiske variable R, G, og B:

R	G	B
100	120	72
105	110	72
110	100	74
130	90	79
150	80	70
150	70	67
180	50	55
200	30	89
120	100	75
160	60	64

   Opstil et estimat af kovariansmatricen, beregn spredningerne af de tre stokastiske variable samt korrelationskoefficienterne. Er der grund til at tro at nogle af variablene er korrelerede? Beregn egenværdier og egenvektorer for kovariansmatricen (brug MATLLAB), og beregn Karhuen-Loeve transformationen af de 10 målinger. Substituér, den mindste af egenværdierne med værdien 0, svarende til en reduktion af dimensionaliteten af målingerne fra 3 til 2. Beregn ved invers transformation de herved approksimerende målinger. Beregn afvigelserne fra de oprindelige målinger.

2. Lad to populationer af observationer $\mathcal{O}_1$ hhv. $\mathcal{O}_1$ være givet ved deres middelværdier $\mathbf{m}_1$ hhv. $\mathbf{m}_2$ samt deres kovariansmatricer $\mathbf{C}_1$ hhv. $\mathbf{C}_1$.

$$\mathbf{m}_1 = (4.48, 2.19, 32.77, 12.28) \qquad \mathbf{m}_2 = (17.30, 5.71, 30.01, 3.93)$$

$$\mathbf{C}_1 = \begin{bmatrix} 4.10 & 1.42 & 0.68 & 5.95 \\ 1.42 & 1.03 & 0.82 & 2.44 \\ 0.68 & 0.82 & 0.96 & 4.53 \\ 5.95 & 2.44 & 4.53 & 24.65 \end{bmatrix} \qquad \mathbf{C}_2 = \begin{bmatrix} 9.07 & 6.34 & 4.87 & 1.89 \\ 6.34 & 4.12 & 4.82 & 4.18 \\ 4.87 & 4.82 & 16.76 & 5.78 \\ 1.89 & 4.18 & 5.78 & 8.86 \end{bmatrix}$$

Hvad er Mahalanobis afstanden fra punktet $\mathbf{m}_1$ til observationsklyngen $\mathcal{O}_2$, og hvad er Mahalanobis afstanden fra punktet $\mathbf{m}_2$ til observationsklyngen $\mathcal{O}_1$. Hvad er divergensafstanden mellem de to populationer? Hvad er Bhattacharyya afstanden mellem populationerne. Vink: Brug MATHLAB (eller MAPLE) til til udregningerne.

3. Du bedes vurdére om en af de fire features, hvis statistiske egenskaber er vist i ovenstående opgave, kan udelades uden stort informationstab. Eliminér iterativt en featurevariabel, og genberegn de i ovenstående opgave nævnte afstandsmål. Beskriv uformelt hvorledes afstandsmålene afhænger af dimensionen af featurevektoren.

4. Lad følgende 5 observationer af en to-dimensional featurevektor (a,b) være givet ved: (1,3); (0,1); (2,5); (4,4); og (3,7). Angiv hvilken værdi som korrelationskoefficienten er tættest på:
2) $\rho_{12} \approx 0.915$
1) $\rho_{12} \approx 0.865$
5) $\rho_{12} \approx 0.707$
7) $\rho_{12} \approx 0.667$
3) $\rho_{12} \approx 0.512$
4) $\rho_{12} \approx 0.500$
6) $\rho_{12} \approx 0.209$
8) Ved ikke
Transformér nu featureværdierne ved a' = a+b, og b' = y-x. Angiv for de transformerede værdier (a',b') hvilken værdi som korrelationskoefficienten er tættest på:
1) $\rho'_{12} \approx 0.903$
2) $\rho'_{12} \approx 0.512$
3) $\rho'_{12} \approx 0.500$
4) $\rho'_{12} \approx 0.475$
5) $\rho'_{12} \approx 0.447$
6) $\rho'_{12} \approx 0.412$
7) $\rho'_{12} \approx 0.209$
8) Ved ikke

Kapitel 3

Lineær Regression

Metoder til håndtering af lineære ligningssystemer er vel nok de mest omtalte og veludviklede i matematikken såvel som i diverse anvendelser. Årsagen hertil er dels, at matematikken bag lineære systemer er klassisk og veludviklet, at der findes meget effektive beregningsmetoder til løsning af lineære ligningssystemer, og ikke mindst, at mange praktiske problemer lader sig modellere godt ved hjælp af lineære ligninger.

Hovedformålet med kapitlet er at introducere til *Mindste kvadraters metode* til analyse af overbestemte lineære ligningssystemer, samt de fordele og ulemper som denne metode tilbyder. Det vises at metoden i visse tilfælde er ustabil, og en række alternative (mere robuste) metoder skitseres. Mindste kvadraters metode er klassisk og et grundelement i mange mere avancerede metoder til dataanalyse. Vi skal eksempelvis søge svar på spørgsmål som: Hvordan tilpasses en lineær model til en række observationer, og hvornår er tilpasningen et maksimum likelihood estimat? Hvordan kan fejlindtastninger eller andre sporadiske fejl detekteres? Hvor mange grove fejl kan accepteres før estimationen fejler? Er det muligt at estimere parametrene selv om op mod halvdelen af observationerne er fejlmålinger?

Et typisk problem i dataanalyse er at beskrive afhængigheden af en *observerbar variabel* y som funktion af et sæt af n *modelvariable* x_j, hvor $j = 1 \dots n-1$. I resten af dette kapitel skal vi antage, at afhængigheden er lineær, dvs. at afhængigheden kan skrives:

$$y = a_0 + \sum_{j=1}^{n-1} a_j x_j = a_0 + a_1 x_1 + a_2 x_2 + \dots + a_{n-1} x_{n-1} = \mathbf{x}_j \mathbf{a} \tag{3.1}$$

hvor $\mathbf{x}_j = (1, x_1, x_2, \dots, x_{n-1})$, og hvor $\mathbf{a} = (a_0, a_1, \dots, a_{n-1})^t$. Parameteren a_j beskriver med hvor stor vægt modelvariablen x_j påvirker observationen y. Ved brug af lineære modeller som den ovenstående er det vigtigt at huske, at lighedstegnet kun udtrykker en formodning. Sagt med andre ord er venstresiden kendt (målt), mens højresiden er en hypotese. Under visse omstændigheder er det muligt, at teste hvorvidt hypotesen er troværdig eller ej. I resten af kapitlet antages at modellen er korrekt, dvs. at y kan modelleres ved en lineær funktion i de $n-1$ modelvariable x_j, samt konstanten 1.

Ved *estimation* af en lineær model ønsker vi at bestemme de n parametre a_j. Denne estimation kaldes også for *lineær regression*. Det er klart, at et krav til enhver estimation er, at antallet af målinger (samhørende værdier af modelvariable og den observerbare variabel) mindst er lig n. Det er naturligvis også et krav, at værdierne af de valgte modelvariable er tilgængelige. Antag derfor at der til hver observation y_i af den datagenererende proces er kendt værdierne af de n modelvariable x_j. Vi kan nu opskrive et system af lineære ligninger:

$$\begin{aligned}
y_1 &= a_0 + a_1 x_{11} + a_2 x_{12} + \ldots + a_{n-1} x_{1(n-1)} \\
y_2 &= a_0 + a_1 x_{21} + a_2 x_{22} + \ldots + a_{n-1} x_{2(n-1)} \\
\vdots \quad & \quad \vdots \qquad\qquad\quad \vdots \qquad\qquad\qquad \vdots \\
y_k &= a_0 + a_1 x_{k1} + a_2 x_{k2} + \ldots + a_{n-1} x_{k(n-1)}
\end{aligned} \tag{3.2}$$

hvor k er antallet af observationer. Lad $\mathbf{X} = (x_{ij})$ være matricen bestående af de $k \times n$ modelvariable. Matricen $\mathbf{X}$ kaldes ofte for *design-matricen*, fordi den specificerer et forsøg bestående af k delforsøg, der hver er bestemt ved angivelse af værdien af de n modelvariable. Lad endvidere $\mathbf{y}$ angive søjlevektoren bestående af de k observerede værdier. Ligningssystemet kan nu opskrives på matrix form:

$$\mathbf{y} = \mathbf{X}\mathbf{a} \tag{3.3}$$

I de fleste praktiske anvendelser er det ikke på forhånd givet, at den opstillede model er eksakt. Tværtimod betragtes modellen som en hypotese, som vi vil forsøge at tilpasse til vores observationer. Dette kan formuleres matematisk ved til tilføje et fejlled $\mathbf{r}$, der modellerer afvigelsen mellem observationsvektoren og modelprediktionen:

$$\mathbf{r} = \mathbf{y} - \mathbf{X}\mathbf{a} \tag{3.4}$$

Afvigelsen $r_i = \mathbf{x}_{i.}\mathbf{a}$ af kaldes for *residualet* for den i'te måling. Vektoren $\mathbf{r}$ kaldes residualvektoren. r_i kan således opfattes som et korrektionsled, der beskriver den del af den observerede værdi y_i, som modellen ikke kan forklare. Hvis modellen vides at være korrekt, kan r_i alternativt opfattes som en målefejl på observationen y_i. I begge tilfælde ønsker man at bestemme de parametre a_i, der opfylder ligningen på en sådan måde, at fejlleddet er mindst muligt. Da $\mathbf{r}$ er en vektor, er det derfor nødvendigt at specificere den norm, hvorunder størrelsen af fejlen skal betragtes. Typisk betragtes den euclidiske norm.

3.1 Mindste Kvadraters Metode

I mindste kvadrates metode søger man finde det sæt af parametre a_i der minimerer kvadratet på modelafvigelsen (målefejlen), idet den euclidiske norm benyttes. Dette svarer til minimering af residualvektorens længde. Denne længde er givet ved:

$$\begin{aligned}
r^2 &= \mathbf{r}^t\mathbf{r} = ||\mathbf{y} - \mathbf{X}\mathbf{a}||^2 \\
&= (\mathbf{y} - \mathbf{X}\mathbf{a})^t(\mathbf{y} - \mathbf{X}\mathbf{a}) \\
&= \mathbf{y}^t\mathbf{y} - \mathbf{y}^t\mathbf{X}\mathbf{a} - \mathbf{a}^t\mathbf{X}^t\mathbf{y} + \mathbf{a}^t\mathbf{X}^t\mathbf{X}\mathbf{a} \\
&= \mathbf{y}^t\mathbf{y} - 2\mathbf{y}^t\mathbf{X}\mathbf{a} + \mathbf{a}^t\mathbf{C}\mathbf{a}
\end{aligned} \tag{3.5}$$

hvor $\mathbf{C} = \mathbf{X}^t\mathbf{X}$, dvs. at elementerne $c_{ij} = \mathbf{x}^{\mathbf{t}}_{\mathbf{.i}}\mathbf{x}_{\mathbf{.j}}$ i $\mathbf{C}$ er gived ved vektorproduktet mellem den i'te og den j'te søjle i $\mathbf{X}$. Heraf følger at $\mathbf{C}$ er en symmetrisk $n \times n$-matrice.

Hvis søjlerne i $\mathbf{X}$ er lineært uafhængige, dvs. hvis der for alle $\mathbf{a} \neq \mathbf{0}$ gælder at $\mathbf{Xa} \neq \mathbf{0}$, da kaldes $\mathbf{C}$ for *positiv definit*. I dette tilfælde gælder oplagt at $\mathbf{a^t Ca} > \mathbf{0}$. Da søjlerne i $\mathbf{X}$ er lineært uafhængige vil $\mathbf{C}$ være ikke-singulær, og vil derfor være invertibel. Lad nu $\mathbf{b} = \mathbf{X}^t\mathbf{y}$, og betragt ligningssystemet:

$$\mathbf{b} \;=\; \mathbf{C}\mathbf{a}^* \;=\; \mathbf{X}^t\mathbf{X}\mathbf{a}^* \tag{3.6}$$

Ifølge det ovenstående har dette kvadratiske system, der kaldes for *normalligningen*, en entydig løsning $\mathbf{a}^*$. Vi skal nu vise, at denne løsning minimerer kvadratresidualet i ligning (3.5). Lad $r^2(\mathbf{a}^*)$ angive værdien af kvadratresidualet i ligning (3.5) når den entydige løsning $\mathbf{a}^*$ benyttes. Lad $\mathbf{h}$ være en vilkårlig vektor i $\mathcal{R}^n$. Da gælder:

$$\begin{aligned} r^2(\mathbf{a}^* + \mathbf{h}) \;&=\; \mathbf{y}^t\mathbf{y} \;-\; 2\mathbf{y}^t\mathbf{X}(\mathbf{a}^* + \mathbf{h}) \;+\; (\mathbf{a}^* + \mathbf{h})^t\mathbf{C}(\mathbf{a}^* + \mathbf{h}) \\ &=\; \mathbf{y}^t\mathbf{y} \;-\; \mathbf{b}^t\mathbf{a}^* \;+\; \mathbf{h}^t\mathbf{C}\mathbf{h} \end{aligned} \tag{3.7}$$

I ligning (3.7) er kun det sidste led afhængig af $\mathbf{h}$. Dette led er endvidere strengt positiv for $\mathbf{h} \neq \mathbf{0}$, da $\mathbf{C}$ er positiv definit. Derfor antager r^2 sin minimale værdi for $\mathbf{h} = \mathbf{0}$. Anderledes sagt stemmer mindste kvadraters løsning til det overbestemte ligningssystem (3.4) overens med løsningen til normalligningen (3.6). Hvis $\mathbf{C}$ er positiv definit er den invertibel. Mindste kvadraters løsning til det oprindelige ligningssystem kan derfor udtrykkes ved:

$$\hat{\mathbf{a}} \;=\; (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} \tag{3.8}$$

Matricen $(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$ kaldes den *pseudoinverse*. Årsagen til at løsningen er "hattet" er, at denne er et estimat, og *ikke* en eksakt løsning (til det oprindelige system (3.3) der jo er overbestemt og derfor ingen løsning har). Som vi skal se senere findes der andre estimatorer (end mindste kvadraters metode), som vil give andledning til andre estimater. Som beskrevet senere indtager mindste kvadraters metode imidlertid en særlig position.

Eksempel 4-1
Antag at vi ønsker at modellere et sæt af k data (x, y) ved et tredieegrads polynomium i x. Modellen er således:

$$y_i \;=\; a_0 \;+\; a_1x_i \;+\; a_2x_i^2 \;+\; a_3x_i^3 \tag{3.9}$$

Hvis $k > 4$ vil ligningssystemet være overbestemt. Designmatricen $\mathbf{X}$ bliver:

$$\mathbf{X} \;=\; \begin{pmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^2 & x_2^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_k & x_k^2 & x_k^3 \end{pmatrix} \tag{3.10}$$

Matricen $\mathbf{X}$ kaldes (i dette tilfælde) for en vandermonde-matrice. Da søjlerne i $\mathbf{X}$-matricen er lineært uafhængige kan vi finde et estimat til løsningen ved ligning (3.8).

Eksemplet illustrerer at kravet til anvendelse af mindste kvadraters metode er, at modellen er lineær i model**parametrene**, ikke i model**variablene**. Eksemplet illustrerer imidlertid også et væsenligt problem. Hvis visse af x-værdierne er $\gg 1$ vil potensopløftningerne bevirke at de tilsvarende rækker i matricen vil få dominerende indflydelse på resultatet. Tilsvarende vil x-værdier $\ll 1$ stort set ingen indflydelse få. Vandermonde-matricer er generelt meget dårligt konditionerede, og skal benyttes med varsomhed selv når graden af det approksimerende polynomium er lille. En lille "trick", der i et vist omfang kan afhjælpe problemet, er at skalere hver søjle i $\mathbf{X}$-matricen, således at alle elementer i denne får omtrentlig samme størrelsesorden. Ligger x-værdierne i intervallet [1:10], og skaleres de fire søjler med faktorene 10^0, 10^{-1}, 10^{-2}, og 10^{-3}, svarer dette til at "løse" systemet for parametrene a_0, $10a_1$, 10^2a_2, og 10^3a_3.

Eksempel 4-1 slut

3.1.1 Mindste kvadraters metode i normalfordelt støj

Antag at den opstillede lineære model er korrekt, at antallet af observationer er større end antal ubekendte, og at søjlerne i design-matricen er lineært uafhængige. Antag videre at de observerede værdier er behæftet en målefejl (støj) , og at disse fejl er ukorrelerede, dvs. at støjen i to målinger er uafhængige. Antag yderligere at støjen er normalfordelt med middelværdi 0 og med samme konstante spredning σ, dvs. at støjen er stationær. Sandsynligheden for observation af en given støjværdi r_i for den i'te måling er da givet ved:

$$p(r_i) \;=\; \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{y_i - \mathbf{x}_{i.}\mathbf{a}}{\sigma})^2} \tag{3.11}$$

Da støjen på hver af de k målinger er uafhængige er sandsynligheden for den samlede fejl lig produktet af de enkelte sandsynligheder.

$$p(\mathbf{r}) \;=\; \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^k \prod_{i=1}^{k} e^{-\frac{1}{2}(\frac{y_i - \mathbf{x}_{i.}\mathbf{a}}{\sigma})^2} \tag{3.12}$$

Lad $\mathcal{O}$ betegne mængden af observationer $(\mathbf{x}_{i.}, y_i)$. En estimator siges at give et *maksimum likelihood estimat* , hvis den bestemmer det mest sandsynlige sæt af parametre, der kan forklare de givne observationer. Dette betyder at estimatoren skal vælge det sæt af parametre, der maksimere sandsynligheden $p(\mathcal{O}|\mathbf{a})$. Givet kendskab til parametervektoren $\mathbf{a}$ kan vi udregne værdien af residualvektoren $\mathbf{r}$. Da vi har antaget, at støjen, dvs. værdierne r_i, er normalfordelte med middelværdi 0, og samme konstante spredning σ, kender vi $p(\mathcal{O}|\mathbf{a})$. Maksimum likelihood estimatet for den lineære regression $\mathbf{y} = \mathbf{X}\mathbf{a}$ er således givet ved maksimumspunktet for funktionen:

$$f(\mathbf{a}) \;=\; p(\mathbf{r}) \tag{3.13}$$

Da logaritmefunktionen er monotont voksende ses let at maksimum likelihood estimatet også er givet ved minimumspunktet for udtrykket:

$$k\ \log(\sqrt{2\pi}\sigma) \;+\; \sum_{i=1}^{k} \frac{1}{2}(\frac{y_i - \mathbf{x}_{i.}\mathbf{a}}{\sigma})^2 \tag{3.14}$$

Idet k og σ er konstanter, og residualet er defineret ved $r_i = y_i - \mathbf{x}_{i.}\mathbf{a}$, ses at maksimum likelihood estimatet opnås for den estimator, der minimerer summen af kvadratresidualerne, dvs:

$$\min \sum_{i=1}^{k} \rho(\mathbf{r}) \;=\; \min \sum_{i=1}^{k} r_i^2 \tag{3.15}$$

Ifølge det foregående er dette netop, hvad mindste kvadraters metode gør. Vi har således vist at denne metode er optimal (i maksimum likelihood forstand) hvis støjen er normalfordelt.

Det er let at vise (se opgaveafsnittet) at summen af residualværdierne er lig 0. Det følger direkte at mindste kvadraters metode minimerer variansen af residualerne.

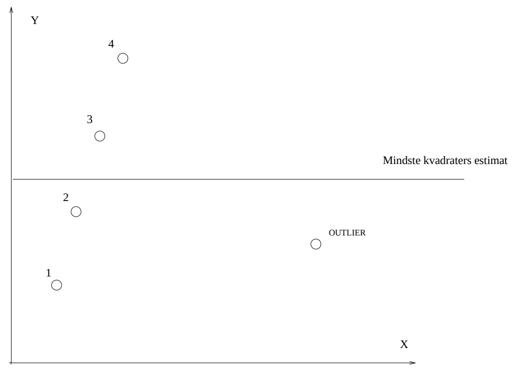
Som vi skal diskutere senere er fejleddet r_i ikke altid normalfordelt med middelværdi 0 og konstant spredning. I visse tilfælde kan støjen bedre modelleres ved en eksponentialfordeling, ved en ligefordeling, eller ved kombinationer af disse. I disse tilfælde vil mindste kvadraters metode ikke (garanteret) give det mest sandsynlige sæt af parameterværdier. Det kan faktisk vises (se senere) at blot én måling er behæftet med en tilstrækkelig stor støjværdi, da kan mindste kvadraters metode resultere i vilkårligt forkerte parameterværdier.

Årsagen til, at mindste kvadraters metode, på trods af ovennævnte forhold, er så populær, er dels, at der findes meget effektive beregningsmetoder til denne metode, dels *De store tal lov* fra sandsynlighedsregningen. Denne siger, at hvis fejlen på en given observation kan beskrives som summen af uendelig mange små fejled, da vil fejlen være normalfordelt, uanset hvilken fordeling hvert af de bidragende fejled måtte have (det er dog et krav at disse fordelinger er ens). I de situationer hvor det er rimeligt at antage, at støjen på en måling kan beskrives som sådan en sum, er mindste kvadraters metode derfor optimal.

3.2 Outliers, Robuste estimatorer

En *Outlier* er en måling $(\mathbf{x}, y)$, der ikke stammer fra den proces, som vi ønsker at modellere. Data, der stammer fra den proces, som vi ønsker at modellere, kaldes *inliers*. Outliers stammer typisk fra fejlaflæsninger (fejlindtastninger) eller fra fejl opstået på et tidligere trin i dataanalysen. I visse tilfælde kan komplicerede processer opføre sig således at de for det meste ligner en lineær proces, men under visse omstændigheder producerer data, der ikke kan forklares så simpelt. Hvis vi kun ønsker at modellere den lineære komponent af processen kan anormalierne bedst karakteriseres som outliers.

Karakteristisk for outliers er, at de optræder sporadisk, og at der ingen sammenhæng er mellem disse og processen som vi ønsker at modellere. Ofte vil outliers give anledning til store residualer, og disse kan betragtes som trukket fra en ligefordeling. Således kan den samlede fordeling af residualerne ofte beskrives godt ved en vægtet sum af en normalfordeling of en ligefordeling. I dette tilfælde er mindste kvadraters metode ikke længere optimal. Betragt tilfældet illustreret i Figur 3.1, hvor 4 punkter fra en lineær model samt en outlier er illustreret. Ved brug af mindste kvadraters metode opnås et fit, der er langt fra den korrekte løsning.



Figur 3.1: Mindste kvadraters løsning på 5 observationer, hvoraf én er en outlier.

Ved at flytte på positionen af outlierpunktet er det let at overbevise sig om, at mindste kvadraters metode kan frembringe ethvert regressionsresultat, hvis blot én outlier er til stede. Dette er yderligere uafhængigt af hvor mange observationer, der iøvrigt er til rådighed, og hvor godt disse stemmer overens med modellen.

Eksempel 3-2
Antag at punkterne (x, y) for de 5 observationer i Figur 3.1 er givet ved $(1,2)$, $(2,5)$, $(3,8)$, $(4,11)$ og $(10,4)$, hvor sidstnævnte er en outlier. Ved fit af en ret linie $\beta + \alpha x$ til disse data er vetoren $\mathbf{y}$ lig $(2,5,8,11,4)^t$, og designmatricen $\mathbf{X}$ givet ved:

$$\mathbf{X} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 10 \end{pmatrix} \quad \mathbf{X}^t\mathbf{X} = \begin{pmatrix} \sum 1 & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} = \begin{pmatrix} 5 & 20 \\ 20 & 130 \end{pmatrix} \tag{3.16}$$

Da determinanten af $\mathbf{X}^t\mathbf{X} = 250$ får vi, at den pseudoinverse matrice bliver:

$$(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t = \frac{1}{250}\begin{pmatrix} 110 & 90 & 70 & 50 & -70 \\ -15 & -10 & -5 & 0 & 30 \end{pmatrix} \tag{3.17}$$

Ved at multiplicere denne matrice med vektoren $\mathbf{y}$ fås at $\alpha = 0$, og $\beta = 6$, som illustreret på figuren. Residualvektoren $\mathbf{r}$ bliver (4, 1, -2, -5, 2).

Eksempel 3-2 slut

Estimatorer kan karakteriseres ved deres *robusthed* , dvs. deres evne til at give rimelige resultater når der optræder outliers blandt observationerne. *Nedbrudspunktet* for en estimator er uformelt defineret ved den mindste andel af outliers blandt observationerne, således at et vilkårligt dårligt estimat kan frembringes. Mere formelt, lad k være lig antal observationer, $\mathcal{Z}$ være mængden af observationer:

$$\mathcal{Z} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_k, y_k)\} \tag{3.18}$$

Lad $T(\mathcal{Z})$ være en estimator, og lad $\mathcal{Z}_m$ være defineret ud fra $\mathcal{Z}$ ved at erstatte m observationer med vilkårlige værdier (outliers). Lad $B(m, \mathcal{Z})$ være den maksimale afvigelse (bias) der kan frembringes ved erstatningen af de m observationer med outliers:

$$B(m, \mathcal{Z}) \;=\; \sup_{\mathcal{Z}_m} ||T(\mathcal{Z}_m) - T(\mathcal{Z})|| \tag{3.19}$$

hvor supremum tages over alle mængder $\mathcal{Z}_m$. Hvis $B(m, \mathcal{Z}) = \infty$ kan m outliers altså have en vilkårlig stor effekt på estimatet. Estimatoren siges at bryde sammen. Nedbrudspunktet for T defineres som:

$$N_T \;=\; \min\left\{\frac{m}{k} \mid B(m, \mathcal{Z}) = \infty\right\} \tag{3.20}$$

Dvs., som den mindste andel af outliers, der kan bevirke at estimatet er vilkårligt langt fra $T(\mathcal{Z})$.

For mindste kvadraters metode er nedbrudspunktet lig $1/k$. Der findes estimatorer med nedbrudspunkt så højt som 0.5 (halvdelen af observationer kan være outliers), samt estimatorer hvor nedbrudspunktet afhænger af observationerne (og kan være større end 0.5). Appendix D indeholder en kort beskrivelse af nogle sådanne metoder. Her skal vi undersøge nogle mere klassiske metoder til identifikation af outliers.

Outliers har typisk meget afvigende koordinater (model parameter værdier eller værdier af den observerbare variabel) i forhold til inliers. Det modsatte argument, at afvigende koordinater indikerer en outlier er mere tvivlsomt, fordi en observation, der både afviger mht. $\mathbf{x}$ og y meget vel kan være en inlier. En klassisk metode til identifikation af outliers er at separere de to tilfælde: Først identificeres punkter med afvigende $\mathbf{x}$-koordinater, derefter punkter med afvigende y-koordinater.

Da matricen $\mathbf{X}$ indeholder al information om $\mathbf{x}$-koordinaterne, er det oplagt at undersøge hvor meget hver række i $\mathbf{X}$ matricen bidrager til estimationsresultatet. Forskellen (residualet) mellem den observerede vektor $\mathbf{y}$ og vektoren $\hat{\mathbf{y}}$ beregnet ved brug af mindste kvadraters estimat er:

$$\mathbf{r} \;=\; \mathbf{y} - \hat{\mathbf{y}} \;=\; \mathbf{y} - \mathbf{X}\hat{\mathbf{a}} \;=\; \mathbf{y} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} \;=\; (\mathbf{I} - \mathbf{H})\mathbf{y} \tag{3.21}$$

Matricen $\mathbf{H} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$, der har dimension $k \times k$, hvor k er antal observationer, har en lang række egenskaber: $\mathbf{H}$ er idempotent og symmetrisk, dvs. $\mathbf{HH} = \mathbf{H}$, og $\mathbf{H}^t = \mathbf{H}$. Der gælder at *trace* $\mathbf{H}$ = *rank* $\mathbf{H} = n$, hvor n er antal regressionsparametre. Kvadratresidualet r^2 er lig $\mathbf{y}^t(\mathbf{I} - \mathbf{H})\mathbf{y}$. Endelig kan det vises at diagonalelementet h_{ii} er lig $\sum_{j=1}^{k} h_{ij}^2$.

Af $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$ følger at elementet h_{ij} beskriver virkningen af den j'te observation på $\hat{y}_i$. Diagonalelementet h_{ii} beskriver således hvorledes den i'te observation indvirker på sin egen estimation. Ifølge de ovennævnte egenskaber er gennemsnittet af h_{ii} lig n/k, og der gælder at $0 \leq h_{ii} \leq 1$. Da $h_{ii} = h_{ii}^2 + \sum_{i \neq j} h_{ij}^2$ ses, at hvis $h_{ii} = 1$ da vil $h_{ij} = 0$ for $j \neq i$. Dette betyder, at hvis h_{ii} er stor, vil den i'te observation være hovedansvarlig for estimationsværdien $\hat{y}_i$. Ideen er derfor, at klassificere de observationer i, hvor h_{ii} er væsentlig større end n/k, som potentielle outliers.

Eksempel 3-3
I eksempel 3-1 er $n = 2$ og $k = 5$. Vi vil klassificere observationer med $h_{ii} > 2n/k = 0.8$ som potentielle outliers. Matricen $\mathbf{H}$ beregnes let til:

$$\mathbf{H} = \frac{1}{250} \begin{pmatrix} 95 & 80 & 65 & 50 & -40 \\ 80 & 70 & 60 & 50 & -10 \\ 65 & 60 & 55 & 50 & 20 \\ 50 & 50 & 50 & 50 & 50 \\ -40 & -10 & 20 & 50 & 230 \end{pmatrix} \tag{3.22}$$

Diagonalelementerne er derfor $(0.38, 0.28, 0.22, 0.20, 0.92)$. Det ses at outlierpunktet klart skiller sig ud.

Eksempel 3-3 slut

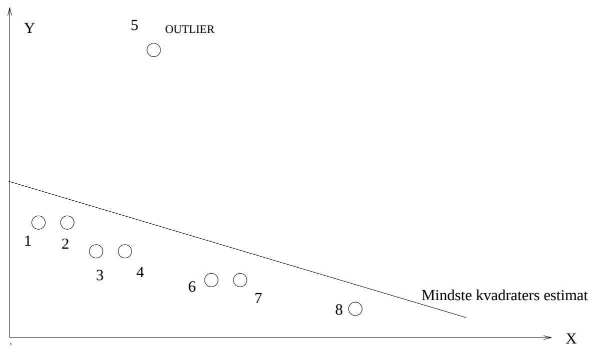
Det skal understreges, at fordi en observation skiller sig ud ved ovenstående analyse, er det på ingen måde sikkert at det er en outlier. Hvis alle observationer ligger perfekt på en ret linie, men et punkt langt fra de øvrige, vil dette punkt (selv om det ikke er en outlier) blive identificeret af ovenstående procedure. Hvis der blandt observationerne findes et større antal af outliers, er det også muligt at ingen af disse vil skille sig ud.

Hvis en observation $(\mathbf{x}, y)$ afviger væsenligt fra de øvrige observationer ved at have en meget afvigende y-værdi, er dette ofte muligt at identificere denne ved at analysere residualvektoren $\mathbf{r}$. Lad σ_r være spredningen af værdierne r_i (opnået ved anvendelse af mindste kvadraters metode). Hvis afvigelsen r_i mellem observationen y_i og modellens forudsigelse $\hat{y}_i$ er stor, da er sandsynligheden for at observationen er en outlier også stor. Hvis det vides at støjen på observationerne er normalfordelt er sandsynligheden givet direkte ved ligning (3.11). Typisk identificeres observationer som potentielle outliers hvis $|r_i| > 2.5\ \sigma_r$.

Eksempel 3-4
Residualerne for fittet beskrevet i eksempel 3-2 var (4, 1, -2, -5, 2). Det ses umiddelbart, at den største numeriske residual ikke optræder for outlieren, men for punkt nummer 4. Spredningen af de 5 værdier er $\sqrt{10}$, svarende til en typisk grænseværdi på $2.5\sqrt{10} \approx 7.9$. Metoden vil altså ikke kunne afsløre outliers af den type, der ses i eksempel 3-2. Det ses også, at hvis den aktuelle outlier fjernes først, vil fittet være perfekt, svarende til en værdi af $\sigma_r = 0$.

Eksempel 3-4 slut

Eksempel 3-5
Antag at punkterne (x, y) for 8 observationer er givet som illustreret i nedenstående figur samt tabel. Bemærk at det femte punkt er en outlier.



Figur 3.2: Mindste kvadraters løsning på 5 observationer, hvoraf én er outlier.

punkt	1	2	3	4	5	6	7	8
x	1	2	3	4	5	7	8	12
y	4	4	3	3	10	2	2	1
r_i	-0.891	-0.593	-1.295	-0.997	6.301	-1.104	-0.806	-0.615
h_{ii}	0.32	0.24	0.18	0.14	0.13	0.16	0.21	0.62

Ved mindste kvadraters fit af en ret linie $\beta + \alpha x$ til disse data fås $\alpha \approx -0.2978$ og $\beta \approx 5.1885$. Residualvektoren r_i er angivet i ovenstående tabel. Spredningen af residualværdierne er $\sigma_r \approx 2.392$, hvorved grænseværdien $2.5\sigma_r$ bliver ca. 5.98. I dette tilfælde ses således at outlieren bliver identificeret. Diagonalelementerne h_{ii} af $\mathbf{H}$-matricen er vist i tabellen. Grænseværdien $2n/k = 0.5$. Det ses, at den sande outlier ikke identificeres ud fra analyse af $\mathbf{H}$-matricen. Derimod identificeres punkt 8 fejlagtigt.

Eksempel 3-5 slut

Et stort problem ved ovennævnte metode er, at σ_r er beregnet på basis af alle observationer (inliers såvel som outliers). Som illustreret i eksempel 4-3 vil selv få outliers kunne bevirke stor ændring af værdien af σ_r. Dette skyldes naturligvis at σ_r^2 beregnes ved $\frac{1}{k}\sum_i r_i^2$. Hvis vi betragter fordelingen af residualværdierne aftager denne meget stærkt (som funktion af $|r_i|$) i normalfordelt støj. Hvis der forekommer outliers vil disse ofte ligge langt fra nulpunktet. Et mere robust mål for σ_r kan derfor opnås hvis halerne af fordelingen ignoreres. Det kan derfor ofte være en fordel at benytte en af de robuste estimatorer, der er bekrevet tidligere i noterne.

3.3 Opgaver

1. Vis ligning (3.7).

2. Beregn den pseudoinverse samt mindste kvadraters løsning for en lineær model bestående af 3 modelparametre, svarende til de 4 observationer (1.0, 3.0, 5.5, 7.8) opnået for de 4 sæt af modelvariable (2.3, 5.1), (6.9, 15.3), (2.7, -1.0), og (7.8, 39.78).

3. Beregn den pseudoinverse samt mindste kvadraters løsning for en lineær model bestående af 3 modelparametre, svarende til de 4 observationer (1.0, 3.0, 5.5, -

7.8) opnået for de 4 sæt af modelvariable (2.3, 5.1), (6.9, -15.3), (2.7, -1.0), og (-7.8, 39.78).

4. Vis at summen af residualerne fremkommet ved mindste kvadraters metode er lig 0. Vink: Brug vektornotation, og vis at summen af elementerne i alle rækker i matricen $\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$ er lig 1.

5. Vis at $\mathbf{H}$-matricen (se ligning (3.21)) er idempotent, samt at kvadratresidualet r^2 er lig $\mathbf{y}^t(\mathbf{I}-\mathbf{H})\mathbf{y}$.

6. Fjern outlieren fra datasættet benyttet i eksempel 3-2, og beregn estimater af σ_r ved de i afsnit 1.2 beskrevne metoder. Hvordan adskiller de sig fra de i eksempel 3-2 anførte værdier.

7. Fjern outlieren fra datasættet benyttet i eksempel 3-5, og beregn estimater af σ_r ved de i afsnit 1.2 beskrevne metoder. Hvordan adskiller de sig fra de i eksempel 3-5 anførte værdier.

8. Kandidatproduktionen på DIKU var i perioden 1984-1993 givet ved

84	85	86	87	88	89	90	91	92	93
18	18	26	31	45	37	28	31	63	46

   Beregn mindste kvadraters fit til kandidatproduktionen under antagelsen at denne, i den aktuelle periode, er lineært relateret til årstallet. Beregn spredningen af residualerne, og undersøg om der findes år (outliers) hvor kandidatproduktionen afviger mere end 2 gange spredningen fra estimationen.

   Fakultetet anslog i en opgørelse den 11. oktober 1993 at DIKU's kandidatproduktion i 1994 ville være 70. Idet den ovenfor udledte simple lineære model antages at holde for 1994 også, skal det vurderes om fakultetets mål for produktionen er realistisk. Under antagelsen at kandidatproduktionen er normalfordelt, med middelvædi og spredning bestemt ved de ovenfor udledte værdier, skal sandsynligheden for at DIKU producerer mindst 70 kandidater i 1994 beregnes.

   Den faktiske kandidatproduktion i 1994 var 38. I 1995 var kandidatproduktionen 46. Gentag beregningen af mindste kvadraters fit, beregn residualværdierne og spredningen af disse. Findes der år, hvor kanditatproduktionen må karakteriseres som usædvanlig (outlier).

9. I et kemisk forsøg har man til temperaturværdierne T målt ledningsevnen l i en substans. De målte data (T, l) var: (1, 1.81); (2, 1.15); (3, 0.65); (4, 0.48); (5, 0.26); (6, 0.14); (7, 0.09); og (8, 0.06). Det vides at den støj, som måleapparatet har tilført data er normalfordelt med middelværdi 0. Optegn de målte data og foreslå en model af disse. Bestem de (i mindste kvadraters forstand) bedst fittende parametre i modellen, og beregn residualerne.

Kapitel 4

Klyngeanalyse

Et automatisk system som genkender mønstre består dels af *sensorer* som omdanner et fysisk signal til en digital vektor, dels af en beslutningsprocedure som genkender vektoren som tilhørende een af flere *klasser*. Ofte reduceres dimensionen af input vektoren hvorved de rå data erstattes af en vektor i en rum af lavere dimension n. Herefter grupperes data i klasser til brug i den senere *klasssifikation*. Vi har tidligere, i kapitel 3, set hvorledes datareduktionen af *featurevektoren* kan foretages ved principalkomponentanalyse.

I dette kapitel beskrives nogle metoder til såkaldt *strukturel mønstergenkendelse* eller klyngeanalyse (nogle gange også kaldt *uovervåget klassifikation*). Disse metoder omfatter bestemmelse af antallet af klasser, hver klasses position i featurerummet, samt variationen af (og korrelationen imellem) de enkelte featureværdier for hver klasse. Denne proces foretages ofte på et sæt af *træningsdata*, hvis korrekte klassifikation er kendt.

Klyngeanalyse benyttes også ved kompression og kodning af data. Dette er eksempelvist nødvendigt ved visning af et RGB- (true color) farvebillede på en 8-bit farvemonitor. En sådan kan vise 256 forskellige farver samtidigt. Dimensionen af featurevektoren er 24, og hver komponent kan antage 2 mulige værdier. Opgaven her er altså at bestemme de 2^8 klasser blandt de 2^{24} mulige forskellige farver. En anden anvendelse er kodning af et gråtone- (8 bit) billede. Dette opdeles ofte i små blokke af f.eks. 4×4 billedelementer. Hver blok er således repræsenteret ved en featurevektor af størrelse $16 \cdot 8 = 128$. Blandt de 2^{128} mulige forskellige delbilleder består klyngeanalysen i at vælge et lille antal (f.eks. 2^8) klasser. Kodningen består i at erstatte hvert delbillede med et indeks til klassen. Denne metode kaldes *vektorkvantificering*. Anvendelse af klyngeanalyse i kodning diskuteres i større detalje i en anden del af kurset.

Klyngeanalyse benyttes ofte i en indledende fase af en dataanalyse. Typisk eksperimenteres med forskellige features. For hvert valg af features benyttes klyngeanalysen sammen med en statistisk karakteristikation af de fundne klasser til vurdering af valget af features (se kapitel 3). Når valget af features er foretaget, og klasserne karakteriseret kan en overvåget (egentlig) klassifikation benyttes til navngivning (se Kapitel 5) af nye observationer.

Hvis antallet af features (dimensionen af observationsvektoren) er lille (1-3), og antallet af klasser er lille (2-10), kan klyngeanalysen ofte foretages interaktivt ved at plotte observationerne, og manualt foretage identifikationen af klynger. Der er udviklet en lang række værktøjer (f.eks. Splus) til hjælp for sådanne eksplorative analyser. Hvis antallet af features er stort eller antallet af klasser er stort er denne metode sjældent en farbar vej frem. I stedet benyttes en algoritme der automatisk kan identificere klyngerne i de forelagte data.

Mønsterrummet (typisk R^n) skal forsynes med en afstand. Det er ingenlunde klart at det skal være den euclidiske afstand. Hvis for eksempel en featureværdi er 0 eller 1, afhængig om en egenskab er tilstede eller ej, er et bedre afstandsmål Hamming afstanden som er antallet af forskellige træk for de to vektorer. Emnesøgning i en database med dokumenter i naturligt sprog forgår fx. ved hjælp af nøgleord som kan opfattes som en reduceret vektor af karakteristiske træk for et dokument.

Da resultatet af en klyngeanalyse er afhængig af valget af features, og af det valgte afstandsmål er klyngeanalyse-processen lidt af en kunst. Det er sjældent muligt at vælge optimale features og afstandsmål i første forsøg. Klyngeanalyse må derfor betragtes som et redskab i en interaktiv proces, med det formål at forstå strukturen i observationerne.

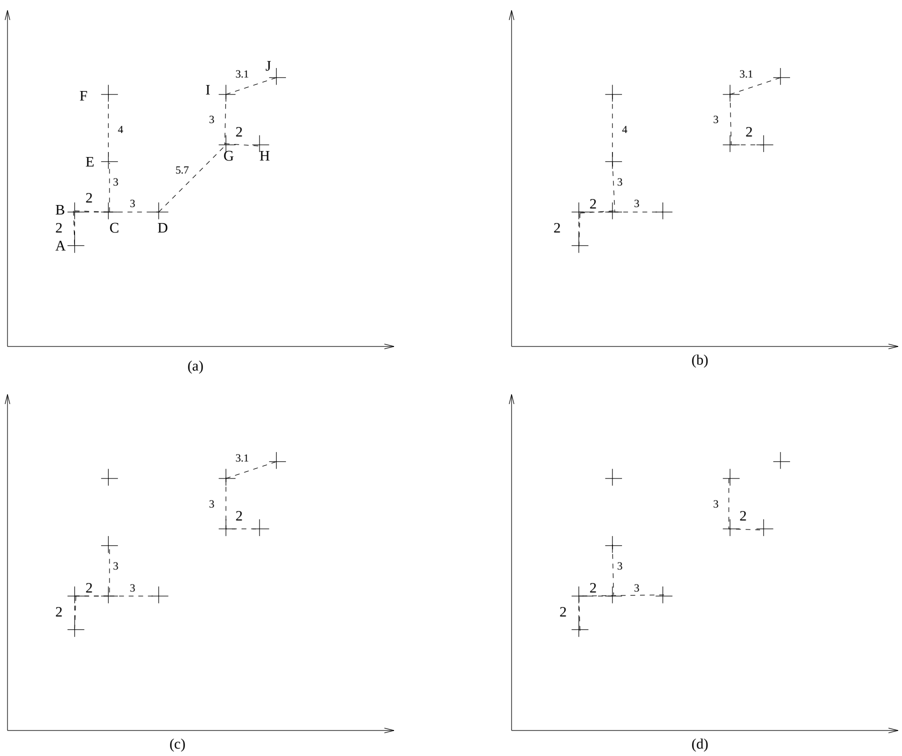
Givet en manuel klassifikation, eller en klassifikation på basis af en klyngeanalyse, er det som regel hensigtsmæssigt at karakterisere de fundne klasser ved en række statistiske mål. Positionen af hver klasse bestemmes ved feature middelværdivektoren for observationerne hørende til hver klasse. Yderligere kan hver klasse karakteriseres ved kovariansmatricen for observationerne. Hver klasse karakteriseres således ved $p + p * (p + 1)/2$ reelle tal, hvor p er antallet af features. Disse størrelser er ikke kendt på forhånd, men må estimeres ud fra de foreliggende observationer. Dette er beskrevet i kapitel 1. Vi skal udnytte denne karakterisering intensivt senere.

4.1 Histogramanalyse

Det første man bør foretage sig med ny data er at tegne deres graf. Er punkterne en mængde af d-dimensionale data skal man altid starte med at tegne fordelingen af hver enkelt komponent for sig. Det svarer til at tegne grafen for projektionen af punktsættet på hver akse. Hvis grafen eller histogrammet, som den ofte kaldes har en tydelig tærskelværdi er det nyttig information til klassificering. Histogrammet kaldes i det tilfælde bimodalt . Den pågældende koordinat er god til at dele datasættet. Tilsvarende bør man om muligt tegne graferne for projektionerne på et plan, det vil sige betragte to koordinater samtidigt. Disse grafer kaldes *scatter plots* eller to dimensionale histogrammer. I en eksplorativ metode *projection pursuit* roteres punktmængden i det d dimensionale rum mens man betragter grafen af projektionen på en plan. På denne måde kan man opdage lineære transformationer af featureværdierne som er nyttige til at opdele data. Metoden kan automatiseres idet man kan søge efter maksimal afvigelse fra en unimodal fordeling, f.eks. en normalfordeling. Dette vil nemlig indikere at der kun foreligger een klasse. Jo mere bimodal, evt. multimodal, fordelingen er, jo bedre er muligheden for datadiskrimination. Programmeringssystemet Splus tillader denne type datanalyse. Man skal anvende programmet *spin*.

4.2 Minimalt udspændende træ

En simpel metode til bestemmelse af klynger i en mænge af n punkter er først at finde et minimalt udspændende træ. Ved trinvis opdeling af træet i deltræer findes et antal sammenhængskomponenter, der hver vil udgøre en klasse. På den måde er punktmængden i første skridt struktureret som en enkelt klasse. Ved at variere en parameter kan vi splitte denne klasse i underklasser og disse igen i underklasser indtil vi har n klasser med hver et element. Den parameter som vi varierer er den største afstand mellem to punkter.



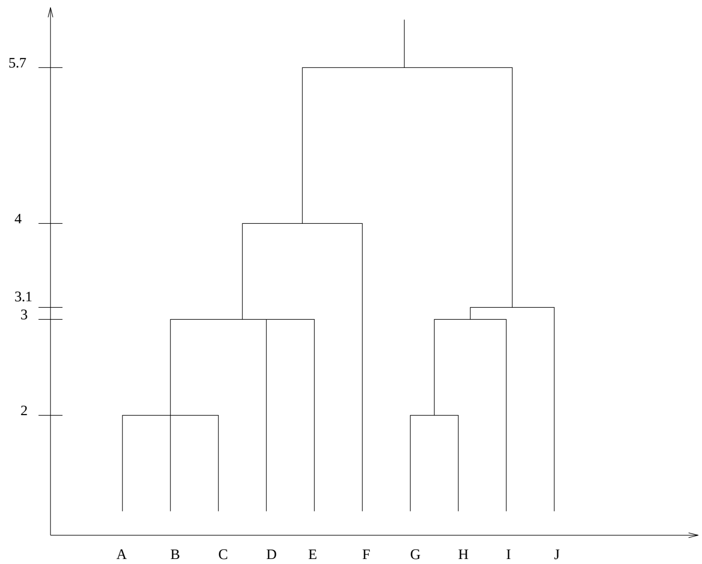
Figur 4.1: Klyngeanalyse ved hjælp af et minimalt udspændende træ. Den mindste afstand mellem to punkter i en klasse formindskes fra (a) til (d).

Eksempel 4.1 Figur 4.1 viser et udspændende træ, hvor de tre længste kanter succesivt er fjernet. Man kan se at den oprindelige ene klasse splitter op i flere efterhånden som kanterne fjernes. Hvor mange kanter skal man nu fjerne for at få en fornuftigt antal klasser? I nedenstående skema har jeg fundet antallet af klasser mellem de kritiske værdier af kantlængden, idet jeg ved en kritisk værdi forstår at der dannes mindst en ny klasse.

Min	Max	Intervallængde	Antallet af klasser
5.7	∞	–	1
4	5.7	1.7	2
3.1	4	.9	3
3	3.1	.1	4
2	3	1	7
0	2	–	10

Et godt og enkelt princip for at finde den rigtige parameterværdi, og dermed at finde antallet af klasser, er at vælge en værdi i det største interval mellem to kritiske værdier af parameteren. På denne måde vælges en struktur (et antal af klasser) som er mest muligt robust over for usikkerhed om parameterens præcise værdi. I ovenstående eksempel forkommer således 2 klasser.

En fordel ved metoden er at klasseinddelingen er hierarkisk: Ved variation af parameteren får man oplysningen om hvorledes en enkelt klasse deler sig i flere underklasser. I klyngeanalyse af for eksempel biologiske data som beskriver forskellige arter kan det være interessant at finde såvel afgrænsing af de enkelte arter som grupperingen af arterne i familier og slægter. Disse forskellige grupperingen kan findes ved at først at lokalisere den mest robuste parameter. Dette kan for eksempel angive de enkelte arter. Dernæst kan den næstbedste værdi bestemmes. Den kan så måske angive geografiske varianter og så videre. En sådan struktur af data angives med et hierarkisk diagram. Figur 4.2 viser en hierarkisk struktur af data fra forrige eksempel.



Figur 4.2: Hierarkisk strukturdiagram. Den vertikale akse viser kritiske parameterværdier.

I stedet for et minimalt udspændende træ kan man samle punkterne i større og større klynger efter **tyngdepunktet** af klasserne. Ved tyngdepunktet for en klasse vil jeg simpelthen forstå et punkt hvis koordinater er punkternes aritmetiske middelværdi. Afstanden mellem to klasser sættes til afstanden mellem deres tyngdepunkter. De to klasser som slås sammen næste gang er de som har kortest indbyrdes afstand. Med denne metode kan vi ligeledes variere afstanden mellem to klasser, bestemme kritiske parameterværdier og danne en hierarkisk strukturering af data.

4.3 Nærmeste nabo.

Såvel i selve opbygningen af klyngecentre som i klassifikationen af en enkelt vektor som tilhørende en af klasserne er nærmeste nabo metoden central. Efter den metode udvælges den klasse, som er nærmeste nabo til den vektor som ønskes klassificeret. Afstanden beregnes fra vektoren til en karakteristisk repræsentant for hver klasse, eventuelt til dens tyngdepunkt. Afstanden kan fortolkes som en usikkerhed i tilhørsforhold. Er for eksempel to klasser næsten lige langt fra den vektor som skal genkendes, kan det være rimeligt midlertidigt at klassificere den som tilhørende begge klasser. Hvis tilsvarende vektoren er meget langt fra alle klasser kan det alt efter anvendelsen være rimeligt foreløbig at lade den forblive uklassificeret. I det mest generelle tilfælde kan proceduren beregne afstanden til samtlige klasser og det kan være svaret.

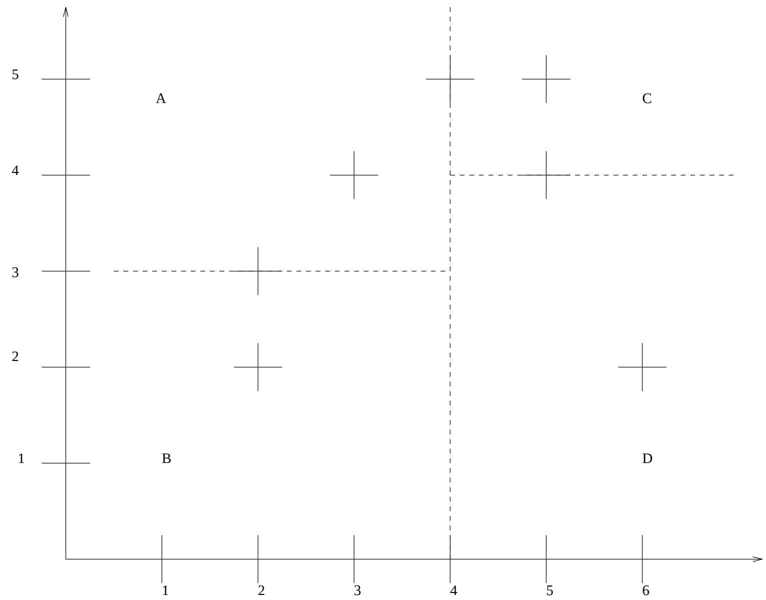
Man kan overveje hvor hurtigt man kan finde den nærmeste nabo. Den naive metode finder afstanden til hver eneste klasse og har derfor en tidskompleksitet på $O(k)$ hvor k er antallet af klasser. Det kan imidlertid vises, at kompleksiteten kan nedbringes til $log(k)$. Dette sker ved opbygning af et k-d-træ.

4.3.1 k-d træer

En hurtig metode til at beregne den nærmeste nabo kan gennemføres ved at opbygge et søgeindex i form af et balanceret binært træ over den givne mængde af punkter n i et d-dimensionalt rum, et såkaldt *k-d træ*.

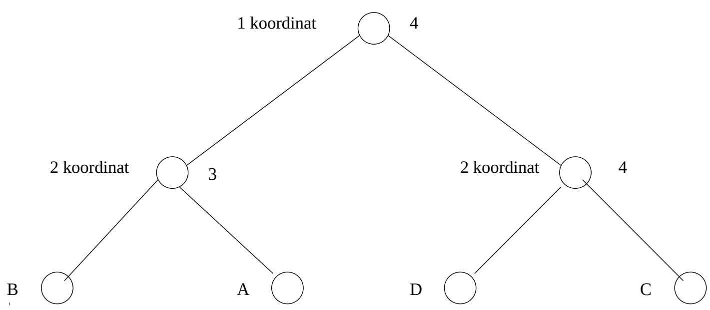
Opbygning af et k-d træ. Mængden af de n punkter underdeles i rekursivt i akseparallelle halvrum. Hver indre knude i træet har tilknyttet et hyperrektangel. Den mængde, som er tilknyttet en knude, er lig med foreningsmængden af de to delmængder som er knyttet til de to undertræer. Et eksempel ses på figur 4.3. Antallet af niveauer i træet er $\lceil \log_2(n) \rceil$.

Ved en forgrening fra en knude til sine to børn opdeles punktmængden for knuden. Opdelingen foregår ved et snit i featurerummet langs en af de d koordinatakser. Konstruktionen forløber fra roden mod bladene, ved at roden først tilknyttes hele punktmængden. I hver indre knude besluttes efter hvilken dimension mængden skal opdeles. Kriteriet er at tage den dimension som udviser den største spredning. Som hurtigt spredningsmål kan man for eksempel tage forskellen på den største og den mindste værdi. På den måde opnår man at underdelingen bliver kompakt. Opdelingen foretages efter valg af en tærskelværdi t_d. Hvis en vektor i delmængden for knuden har den d-te featureværdi mindre end eller lig med t_d tilknyttes den det venstre undertræ, i modsat fald tilknyttes den det højre undertræ. En knude med kun et element tilknyttet har ingen undertræer. t_d vælges som medianen af værdierne. Dette vil sikre at træet bliver balanceret.



Figur 4.3: Et eksempel på en underdeling af en mængde af punkter i forbindelse med konstruktion af et k-d træ. $A = \{(3,4)\}$. $B = \{(2,2),(2,3)\}$. $C = \{(4,5),(5,5)\}$. $D = \{(5,4),(6,2)\}$.

Søgning Når nærmeste nabo til en vektor skal findes, foretages et dybde-først gennemløb af træet. Programmet vedligeholder en global variabel med den hidtil nærmeste nabo. Den initialiseres ved først at søge ned til et blad, idet man i hver knude vælger det undertræ hvis delmængdes d-te feature interval indeholder vektorens d-værdi. I bladet beregnes afstanden fra vektoren til det enlige element. Med denne værdi som radius afgrænses en kugle i featurerummet. Når vi derefter gennemsøger de resterende knuder i træet kan vi undlade afsøgning af en delmængde, og dermed gennemsøgning af et undertræ, hvis dens fællesmængde med kuglen er tom. Hvis vi finder en endnu nærmere nabo formindskes kuglens radius og der skal gennemsøges endnu færre undertræer i det resterende træ. Vi kan sige at algoritmen fungerer efter et *branch and bound* princip . Endvidere kan vi afs-



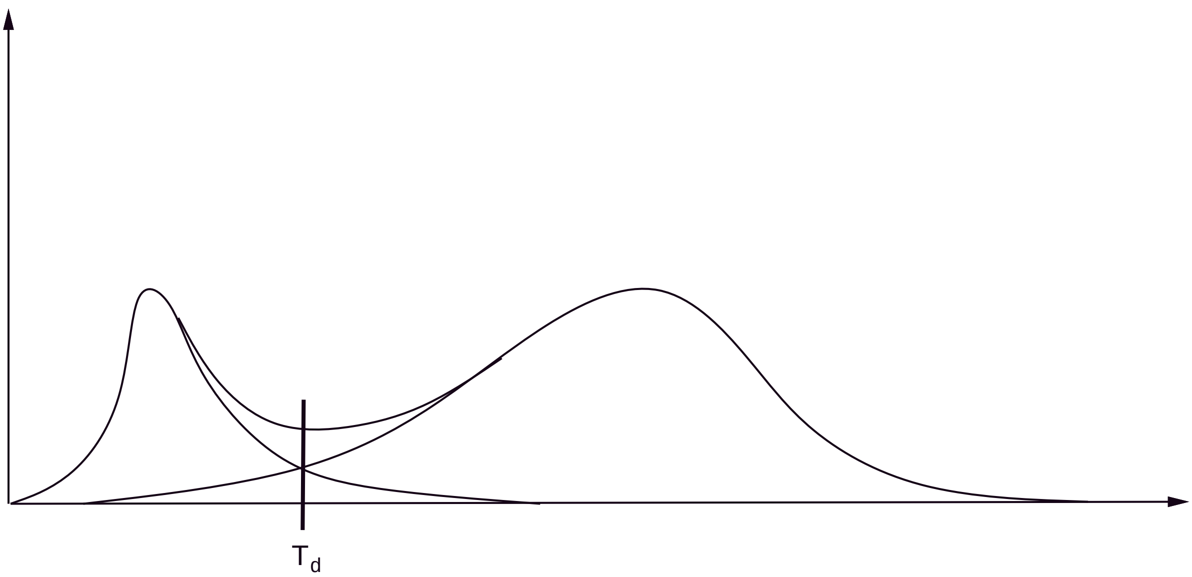
Figur 4.4: I eksemplet er underdelingen ikke ført helt til bunds, idet kun A indeholder et enkelt punkt. Det har vist sig i praksis at der er bedre at have et lille antal, dvs fra 2 til 6 punkter i hvert blad. I et blad undersøges man den mindste afstand fra søgevektoren til samtlige punkter i delmængden.

lutte søgningen fuldstændigt efter behandling af en knude, hvor kuglen ligger helt inden i dens delmængde. Dette skyldes, at vi jo så ved at kuglen har en tom fællesmængde med alle andre delmængder.

Tidsforbrug Konstruktionen af træet kræver i hver knude tid proportional med den betragtede delmængdes størrelse n_i, idet medianen kan findes i tid $O(n_i)$. Derefter gennemløbes knudens delmængde for at danne undertræernes delmængder. Summeret over alle knuder kan vi se at alle n punkter bliver gennemsøgt en gang for hvert niveau i træet. Prisen for at opbygge træeter er derfor $n\lceil log_2 n\rceil$. Det kan vises, at den gennemsnitlige søgetid er proportional med $\log n$, som man kunne håbe.

4.3.2 Kasseopdeling

Et godt estimat af antallet af klynger samt deres position i featurerummet kan fås ved en let modifikation af algoritmen til konstruktion af et $k-d$-træ. Som nævnt foretager denne algoritme opdeling af en punktmængde efter den akse hvorpå projektionen af punktmængden er mest bimodal/multimodal. Hvis der ikke findes nogen feature, hvis tilhørende fordeling er klart bimodal/multimodal, er det nærliggende at stoppe videre opdeling af deltræet. I stedet vil bladet angive en klynge. Positionen af klyngen kan approksimeres ved tyngdepunktet af punkterne hørende til bladet.



Figur 4.5: Skitse af en bimodal fordeling, og de to komponenter af fordelingen. Et bud på en god grænseværdi at opdele populationen ved er angivet ved T_d.

Hvis der findes flere features med tilhørende bimodale projektioner vælges at foretage opdelingen langs aksen for den mest bimodale fordeling. Opdelingen bør **ikke** ske efter medianværdien. En bedre grænseværdi t_d vil være den, der bedst muligt skiller de to delfordelinger, dvs. den værdi hvor summen af de to spredninger er minimal. En approksimation til denne værdi er positionen af den dybeste dal i fordelingen. Bemærk, at dette valg betyder at træet ikke vil blive balanceret. Det skal bemærkes, at det ingenlunde er enkelt at afgøre hvorvidt en fordeling er bimodal eller ej. I visse tilfælde (f.eks. hvis de to antagede komponenter af fordelingen har en enkel parametrisk form, f.eks. normalfordelinger) er det muligt at beregne en optimal værdi af t_d. Det vil imidlertid føre for vidt her, at beskrive sådanne metoder.

4.4 Maximin algoritmen

En anden strategi for bestemmelse af (initielle) klynger er at den indbyrdes afstand mellem disse skal maksimeres. Dette svarer i nogen grad til strategien bag opdeling after det minimalt udspændende træ. *Maximin-algoritmen* er en ofte refereret algoritme af denne type. Først vælges et tilfældigt punkt, som repræsentant for det første klyngecenter z_0. Det andet klyngecenter z_1 bestemmes ved det punkt, der har størst afstand til z_0. Herefter tilføjes nye klyngecentre iterativt indtil en vis betingelse er opfyldt. Lad T være afstanden mellem punkterne z_0 og z_1. Metoden er nu :

1. Beregn for hver af de resterende punkter x afstanden r_{x,z_i} til alle hidtil fundne klyngecentre z_i, og registrer (for hvert x) den mindste af disse afstande r_x^{min}.

2. Find det punkt y, der blandt de resterende punkter x, maksimerer afstanden r_x^{min}, og lad r_0 være denne afstand.

3. Hvis $r_0 \geq \alpha T$ er et nyt klyngecenter fundet. Dette har positionen givet ved y. Lad T være middelværdien af de mindste afstande mellem de nu registrerede klyngecentre. Fortsæt søgningen efter nye klynger som beskrevet under punkt 1).

4. Hvis $r_0 < \alpha T$ er alle klyngecentre fundet. Alle punkter løbes igennem og klassificeres efter nærmeste klyngenabo.

hvor $0 < \alpha < 1$ er en konstant (f.eks. 0.5), der styrer hvor villig algoritmen er til oprettelse af nye klynger. En fordel ved *Maximin-algoritmen* er at den, ud over α, ikke kræver angivelse af nogen parameter. Algoritmen vil (som den baseret på mindste udspændende træer) fungere bedst på relativt kompakte klynger, og vil have en tendens til at opdele meget aflange klynger. Algoritmen er følsom mht. valg af det første klyngecenter. Envidere skal det bemærkes (som for de øvrige metoder) at afstandsbegrebet ikke er specificeret. I mange sammenhænge vil det sædvanlige euclidiske afstandsbegreb være hensigtsmæssigt. Valget af afstandsfunktionen bør dog altid ske ud fra en analyse af de aktuelle data og deres fortolkning (samt ud fra beregningsmæssige hensyn).

Tidskompleksiteten af maximin-algoritmen afhænger af antallet af klynger k i de aktuelle data. Antalet af gange afstanden mellem et punkt og en klyngecenter skal beregnes er (for $k \ll n$) approksimativt lig $n \cdot k$, hvor n er antallet af punkter. Medmindre k er stor vil en repræsentation af klyngerne ved et k-d-træ næppe kunne betale sig. Benyttes et k-d-træ kan kompleksiteten nedbringes til $nlog(k)$.

4.5 K-means algoritmen

I *k-means* algoritmen antages at antallet k af klasser er kendt. Ofte benyttes en anden algoritme, f.eks, maximin-algoritmen, til at bestemme antallet af klynger, samt deres approksimative position i featurerummet. Alternativt kan klyngecentrene vælges tilfældigt, eller så spredt som muligt. Derefter navngives alle observationer efter den klynge, der ligger tættest på observationen. Nye klyngecentre udregnes ved middelværdivektoren af observationerne inden for hver klynge. Derefter itereres navngivningen og genberegningen af klyngecentre sålænge blot ét klyngecenter flytter sig.

1. Lad iterationsnummeret $t = 1$. Vælg k klyngecentre $\mathbf{C}_i^t$, f.eks. ved featurevektorene for de første k af de ialt n observationer.
2. Navngiv hver observation $\mathbf{x}$ med klyngenavnet j hvis der for alle $i \neq j, 1 \leq i \leq k$ gælder at:
$$||\mathbf{x} - \mathbf{C}_j^t|| \; < \; ||\mathbf{x} - \mathbf{C}_i^t||$$
3. Beregn nye klyngecentre ved:
$$\mathbf{C}_i^{t+1} \; = \; \frac{1}{n_i^t} \sum_{\mathbf{x} \in K_i^t} \mathbf{x}$$
hvor K_i^t er klynge nummer i, og hvor n_i^t er antallet af observationer med klyngenavnet i.
4. Lad c være et konvergensmål, f.eks. bestemt ved:
$$c = \sum_{l=1}^{k} ||\mathbf{C}_i^{t+1} \; - \; \mathbf{C}_i^t||$$
Lad $t := t + 1$. Hvis $c = 0$ så stop, ellers fortsæt med punkt 2.

Ovenstående simple metode kan modificeres på en lang række punkter således at nye klyngecentre kan introduceres, således at klynger med meget få "medlemmer" kan nedlægges etc. En ofte refereret metode af denne type er *ISODATA* algoritmen [1,5].

Væsentlig for *k-means* algoritmen er kravet om kendskab til antallet af klyngecentre, valget af initielle klyngecentre samt det afstandsmål hvormed observationer sammenlignes med klyngecentre. En fordel ved metoden er, at den vil kunne fungere fint i situationer hvor observationer fra flere klynger ligger nær hinanden. I disse situationer vil metoder baseret på minimalt udspændende træer eksempelvis fungere dårligt. Beklageligvis findes der intet bevis for konvergens af *k-means* algoritmen. I praksis fungerer metoden imidlertid ofte tilfredsstillende.

Ved hovedløs anvendelse af det euclidiske afstandsmål vil resultatet sjældent svare til det forventede. Dette skyldes at visse featureværdier varierer mere end andre (globalt set), og at visse featureværdier varierer forskelligt afhængig af den korrekte klassifikation. Vi skal ikke her diskutere de forskellige afstandsmål, dette er omtalt i en anden del af kurset.

Tidskompleksiteten af k-means algoritmen afhænger af antallet af iterationer. Den dyre del af algoritmen er klassifikationen (punkt 2 i skitsen). Dette gælder især hvis afstandsmålet er kompliceret. Ved en direkte implementation vil klassifikationen af et enkelt punkt tage en tid, der er proportionalt med antallet k af klynger. Det kan vises at en smartere implementation (ved k-d-træer) kan reducere dette forhold til $log(k)$. Denne implementation er fordelagtig hvis k er stor. For n punkter og t iterationer fås så en tidskompleksitet på $t \cdot n \cdot log(k)$, hvor t er antal iterationer.

4.6 Opgaver

1. Anvend princippet om maksimal robusthed til at finde en klassedeling af punkterne fra eksempel 4.1. Benyt city block afstanden:

$$c((x_1, y_1), (x_2, y_2)) = |x_1 - x_2| + |y_1 - y_2|$$

   i stedet for den euclidiske afstand. Punkterne har koordiinater $A = (3,6), B = (3,8), C = (5,8), D = (8,8), E = (5,11), F = (5,15), G = (12,12), H = (14,12), I = (12,15), J = (15,16)$.

2. Lad $P = \{A, B, C, D, E, F, G, H, I, J\}$ som i forrige opgave. Find den hierarkiske struktur med tyngdepunktsmetoden (se eksempel 4.2).

3. Modificer nærmeste nabo algoritmen (afsnit 4.3.1) til at finde de *to* nærmeste punkter i M.

4. Implementér en klyngeanalyse af én-dimernsionelle data baseret på *k-means*-algoritmen. Lad antallet k af prototyper være en parameter til programmet. Diskutér, implementér, og afprøv nogle forskellige måder hvorpå de initielle prototyper udvælges.

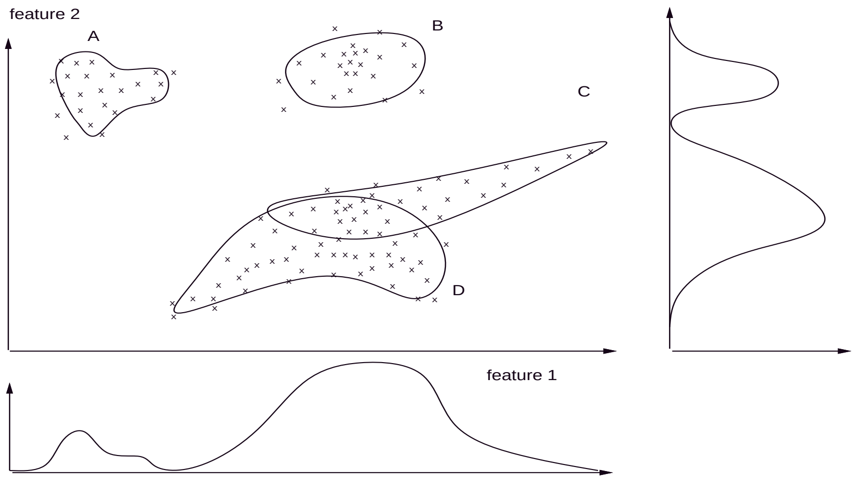
5. Bestem antal og position af klynger ved anvendelse af maximin algoritmen på punkterne angivet i Opgave 1. Lad punktet A være det første tilfældigt valgte klyngecenter, og benyt det euclidiske afstandsmål.

6. Som ovenstående opgave, men med D som det første tilfældigt valgte klyngecenter.

Kapitel 5

Klassifikation

Formålet med en *klassifikation* er at afgøre hvilken *klasse* en given observation stammer fra. Det antages at alle observationer fra samme klasse deler en række fællestræk, der således karakteriserer klassen. Opgaven, at bestemme hensigtsmæssige mål for de fælles karakteristika betegnes *feature ekstraktion* og er diskuteret i et tidligere kapitel. I statistisk klassifikation er opgaven at bestemme en klassifikationsmetode, dvs. en procedure, hvormed en ny observation navngives efter en af de i træningsfasen bestemte klasser. Et stort problem er, at der ofte kan forekomme stor variation af hver feature-værdi inden for hver klasse. Dette er illustreret i nedenstående figur, hvor fire klasser af observationer, hver karakteriseret ved to features er illustreret.



Figur 5.1: Observationer hørende til 4 klasser plottet som funktion af 2 features. Klasse A er velsepareret fra de øvrige. Klasse B og C overlapper som fuktion af feature 1, men ikke som funktion af feature 2. Observationerne hørende til klasse C og D er ikke vel separerede hverken som funktion af feature 1, eller som funktion af feature 2.

I visse heldige situationer kan features vælges således at de separerer klasserne fuldstændigt. Denne situation er imidlertid sjælden. Generelt vil der der findes observationer, der faktisk stammer fra klasse i spredt overalt i featurerummet, og derfor ikke umiddelbart kan skelnes fra observationer stammende fra klasse j. Klassifikationsfejlen er derfor sjældent lig 0.

I visse situationer er antallet klasser kendt på forhånd. Dette er eksempelvis tilfældet i mange industrielle opgaver. Et eksempel er genkendelse af tegn i et OCR-system (eng. Optical Character Recognition). Antag at en tekst er scannet, at de enkelte linier i teksten er separeret og at de enkelte tegn i hver linie er fundet. Ofte er disse repræsenteret ved et binært rasterbillede. De enkelte billedelementer (pixels) har værdi 0 hvis det tilsvarende område på papiret var sværtet, og værdi 1, hvis området var usværtet (hvidt). Hvis de tilladte tegn er dem, der findes på mit tastatur, vil der være ialt 99 klasser (29 små bogstaver, 29 store bogstaver, 10 tal, og 31 specialtegn). Da trykkeprocessen sjældent er ideel, papiret ikke er fuldstændig homogent, og scanneprocessen ikke er fejlfri, vil billederne af tegnene ikke svare fuldstændigt til prototyperne på disse. Opgaven er derfor at bestemme en procedure for bestemmelse af det tegn, der forelagt et billede, ligner bedst.

I andre situationer er hverken antallet af klasser eller fordelingen af observationerne inden for hver klasse kendt. Ved kortlægning af et dyrkningsmøster på basis af multispektrale flyfotos er opgaven at bestemme typerne af afgrøder, hvor store arealer der benyttes til hver sort mv. Da en given afgrødes multispektrale signatur varierer over tid siden såning, med markens egenskaber (herunder jordbundsforhold og hældning i forhold til solen) er det vanskeligt at definere prototyper. Ofte må antallet og karakteriseringen af klasserne bestemmes løbende ud fra de til rådighed værende observationer.

Man skelner mellem *overvåget* klassifikation og *uovervåget* klassifikation. I overvåget klassifikation antages det at et træningssæt af observationer med korrekt klassifikation er givet. Hver klasse er således implicit karakteriseret ved et sæt af observationer. Disse kan benyttes til estimation af fordelingen af observationer som funktion af featurevektoren. I *uovervåget klassifikation* forudsættes intet kenskab til den korrekte klassifikation af observationerne. Antallet af klasser er ofte ukendt. Typisk er formålet med en uovervåget klassifikation netop at bestemme antallet af klasser, samt deres beliggenhed mv. i featurerummet. Denne proces kaldes også for *klyngeanalyse*. Når valget af features er foretaget, og klasserne er fundet og karakteriseret, kan en overvåget (egentlig) klassifikation benyttes til navngivning af nye observationer.

5.1 Feature matching

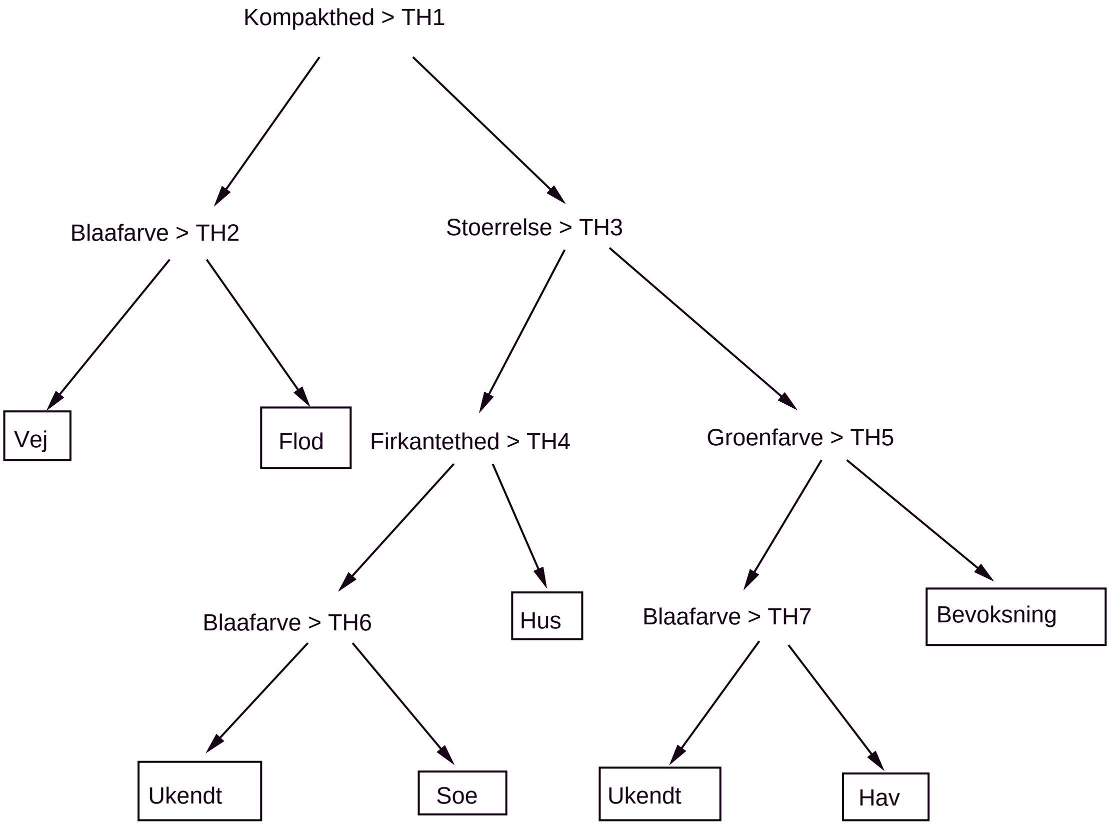
Antag, som eksempel, at vi ønsker at klassificere et billede af et tegn til et alfabet, at tegnet er givet ved et digitalt billede, og at samtlige tegn i alfabetet er givet ved en *prototype* af billedet. En oplagt metode er da, for hvert tegn i alfabetet, at krydskorrelere prototypebilledet for dette med billedet af det observerede tegn. Herved tilknyttes hver mulig fortolkning et godhedsmål (korrelationskoefficienten) for matchet. Observationen klassificeres til det tegn i alfabetet, der opnår størst godhedsmål. Denne metode kaldes *template matching*. Metoden er generelt anvendelig, men ofte langsom, fordi antallet af tegn i alfabetet kan være stort. Et andet problem kan være at godhedsmålet for et match kan være omtrentligt lige stort for et antal prototyper. Valget af vinder kan derfor bero på mængden af støj (forvrængning mv.) af observationen.

For at mindske de beregningsmæssige omkostninger ved klassifikationen er en mulighed at knytte et lille antal features til observationen. Selv om disse kan være beregningskrævende vil der ofte være en samlet beregningsmæssig gevinst fordi vi nu (for hver prototype) kun skal sammenligne featureværdierne. Denne sammanligning kræver fastlæggelse af en norm. Ved direkte brug af den euclidiske norm vil man ofte få uventede resultater, fordi alle features vægtes ens ved denne norm. I praksis varierer visse features mere end andre. Derfor er en normalisering ofte påkrævet. Hertil kommer at visse features ofte betyder mere end andre. Erfaring, evt. opnået gennem maskinel indlæring, kan være afgørende for et velfungerende system. Alternativt kan tænkes at de enkelte feature ekstraktions modeller resulterede i en binær værdi (accept eller reject). Det kan argumenteres at dette blot flytter problemet fra klassifikationsmodulet til ekstraktionsmodulerne. Klassifikationen blever her en simpel optælling af stemmer. Den fortolkning, som flest feature ekstraktionsmoduler stemmer på, vinder. Denne metode er ofte mere robust (pga. kvantificeringen til binære værdier). Ofte vil flere fortolkninger imidlertid få samme stemmetal. Yderligere semantisk information er derfor påkrævet for at kunne afgøre tvivlstilfældene.

Hvert feature ekstraktionsmodul med binær værdi kan betragtes som en test. I situationer, hvor tid er afgørende, er det ofte en god ide, at rangordne de enkelte test i et træ. Således udføres en test kun hvis de foregående test ikke har kunnet afgøre klassetilhørsforholdet. Et sådant træ kaldes et klassifikationstræ . Opbygningen af et klassifikationstræ er vanskelig, bl.a. fordi det ikke er klart hvilke test der bør foretages før andre test. Som ved opbygning af et Hoffman-træ kan det argumenteres, at klasserne for de mest sjældne observationer bør optræde dybt i træet som muligt. Hvis de enkelte test er lige beregningskrævende vil dette minimere den gennemsnitlige klassifikationstid. Ofte er visse test dog meget mere krævende end andre. Disse test, der jo gerne skulle anvendes så sjældent som muligt, og derfor optræde så dybt i træet som muligt, er imidlertid også ofte de mest sigende om klassetilhørsforholdet. Konklussionen er derfor, at klassifikationstræer potentielt giver en stor hastighedsforøgelse, men kræver godt kendskab til de benyttede feature ekstraktionsmetoder, og til den a priori hyppighed af forrekomster af observationer fra de enkelte klasser.

5.2 Beslutningsfunktioner

Det grundlæggende problem i designet af en klassifikationsmetode er at opdele featurerummet i delrum, der hver svarer til en klasse. Forelagt en ny observation består klassifikationen i at bestemme, hvilket delrum featurevektoren tilhører. Opdelingen af featurerummet kan foretages på mange måder. Den kan tage udgangspunkt i en statisktisk karakterisering (middelværdi og kovariansmatrice) af hver klasse, på basis af en klyngeanalyse, eller



Figur 5.2: Eksempel på et simpelt klassifikationstræ for et segment med feature attributterne: Størrelse, kompakthed, firkantethed, og farve.

direkte ud fra et sæt af observationer hvis korrekte klassifikation er kendt. Ofte ønskes at opdelingen er disjunkt og at hele featurerummet dækkes. Opdelingen specificeres som regel ved grænserne mellem nabo-klasser. Ofte antages, at disse grænser har en matematisk simpel form, f.eks. lineære, kvadratiske, eller stykvis lineære.

I det følgende vil vi betegne den i'te klasse med ω_i. En *beslutningsfunktion* $g_{ij}(\mathbf{x})$ er en reel funktion, defineret på featurerummet, med den egenskab at mønsterklasserne i og j befinder sig på hver side af hyperfladen $g_{ij}(\mathbf{x}) = 0$, dvs. således at:

$$\begin{aligned} g_{ij}(\mathbf{x}) > 0 \quad \text{hvis} \quad \mathbf{x} \in \omega_i \\ g_{ij}(\mathbf{x}) < 0 \quad \text{hvis} \quad \mathbf{x} \in \omega_j \end{aligned}$$

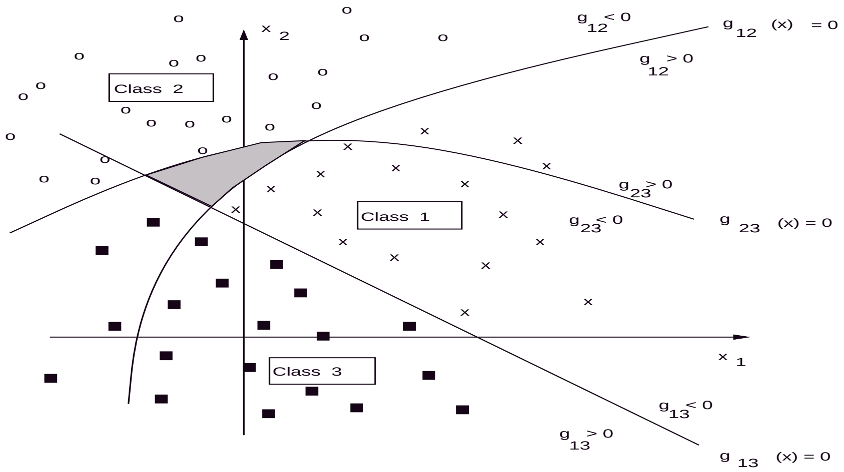
Der gælder at $g_{ij}(\mathbf{x}) = -g_{ji}(\mathbf{x})$. Hvis der ialt er L klasser vil der derfor være $L(L-1)/2$ beslutningsfunktioner. Lad $c(\mathbf{x})$ være en *klassifikationsfunktion*, bestemt ved:

$$c(\mathbf{x}) = i \qquad \text{hvis} \qquad \mathbf{x} \in \omega_i$$

Denne funktion kan jvf. det ovenstående specificeres ved:

$$c(\mathbf{x}) = i \qquad \text{hvis} \qquad g_{ij}(\mathbf{x}) > 0 \quad \forall j \neq i$$

På grænsen mellem to områder i featurerummet vil det ikke være muligt at bestemme en entydig klassifikation. En mulighed er, i dette tilfælde, at vælge klassetilhørsforholdet tilfældigt. Ligeledes vil der ofte være et område hvor det ikke vil være muligt at afgøre klassetilhørsforholdet. Nedenstående figur illustrerer brugen af beslutningsfunktioner for et 3-klasse problem.



Figur 5.3: Beslutningsfunktioner for et 3-klasse problem med 2 features. I de grå område er det ikke muligt at foretage nogen entydig klassifikation.

For at undgå problemerne med områder, der ikke har nogen entydig klassifikation, samt for at reducere antallet af forskellige beslutningsfunktioner antages ofte at disse kan skrives:

$$g_{ij}(\mathbf{x}) = d_i(\mathbf{x}) \; - \; d_j(\mathbf{x})$$

hvor d_i kaldes en *diskriminantfunktion*. Klassifikationsfunktionen bliver nu:

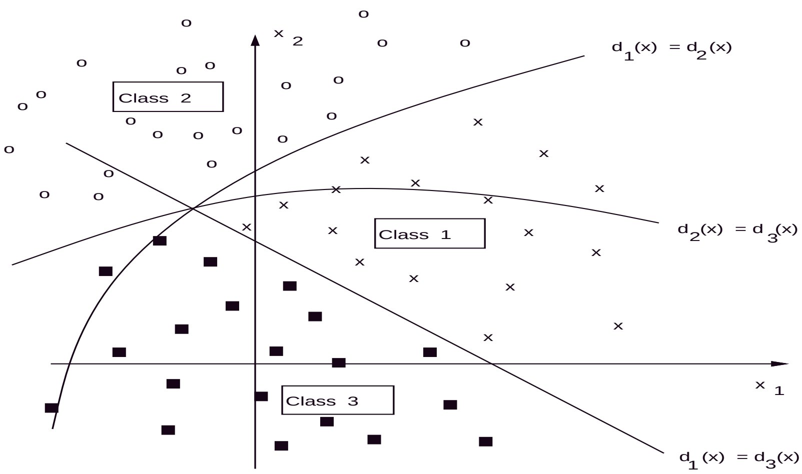
$$c(\mathbf{x}) = i \qquad \text{hvis} \qquad d_i(\mathbf{x}) \; > \; d_j(\mathbf{x}) \qquad \forall j \neq i \tag{5.1}$$

Der findes netop L diskriminantfunktioner. Grænsefunktionerne, bestemt ved $d_i(\mathbf{x}) = d_j(\mathbf{x})$ vil alle skære hinanden i ét fælles punkt. Dette er illustreret i nedenstående figur. Analyser baseret på konstruktion og brug af diskriminantfunktioner kaldes *diskriminant-analyse*.

De mest benyttede diskriminantfunktioner er lineære. Ved lineær diskriminantanalyse i et n-dimensionalt featurerum kan diskriminantfunktionen skrives:

$$\begin{aligned} d_i(\mathbf{x}) &= w_1^i x_1 + w_2^i x_2 + \cdots + w_n^i x_n + w_{n+1} \\ &= \mathbf{w}_i^t \mathbf{x} + w_{n+1} \end{aligned}$$

Hyperplanerne H_{ij}, der udgør grænsefunktionerne, er bestemt ved $d_i(\mathbf{x}) = d_j(\mathbf{x})$. Lineære diskriminantfunktioner kan være gode hvis klasserne er vel separerede. Hvis der findes et sæt af lineære diskriminantfunktioner, der separerer klasserne fuldstændigt kaldes disse for *lineært separable klasser*. Vi skal senere i dette kapitel (samt i kapitel 7) diskutere lineære diskriminantfunktioner i større detalje.



Figur 5.4: Diskriminantfunktioner for et 3-klasse problem med 2 features.

5.2.1 Minimal afstands klassifikatorer

Antag at vi kun kender klassernes position i featurerummet. Denne position kan betragres som en *prototype* på observationerne i klassen. Hvis hver prototype er en god approksimation til enhver observation fra klassen vil en klassifikationsmetode baseret på minimal afstand ofte være effektiv. Klassifikationen er bestemt ved beslutningsreglen (5.1) samt definitionen:

$$d_i(\mathbf{x}) \;=\; -\; dist(\mathbf{x}, \mathbf{x}_i) \tag{5.2}$$

hvor $\mathbf{x_i}$ er prototypen for klasse i, og $dist()$ er en afstandsfunktion. Som beskrevet i en anden del af kurset er valget af afstandsfunktion ikke let. En ofte benyttet afstand er den euclidiske. I dette tilfælde bliver klassifikationsfunktionen:

$$c(\mathbf{x}) = i \qquad \text{hvis} \qquad \sum_{k=1}^{n}(x_k - x_k^i)^2 \;<\; \sum_{k=1}^{n}(x_k - x_k^j)^2 \quad \forall j \neq i$$

hvor n er dimensionen af $\mathbf{x}$. Brugen af det euclidiske afstandsmål forudsætter som regel at hver af featureværdierne er skaleret således at de (totalt set) varierer lige meget. I modsat fald vil visse features ofte få dominerende indflydelse på afstandsmålet. Ved udregning ses at ovenstående klassifikationsfunktion svarer til den lineære diskriminantfunktion:

$$d_i(\mathbf{x}) \;=\; \sum_{k=1}^{n} x_k x_k^i \;-\; \frac{1}{2}\sum_{k=1}^{n}(x_k^i)^2 \tag{5.3}$$

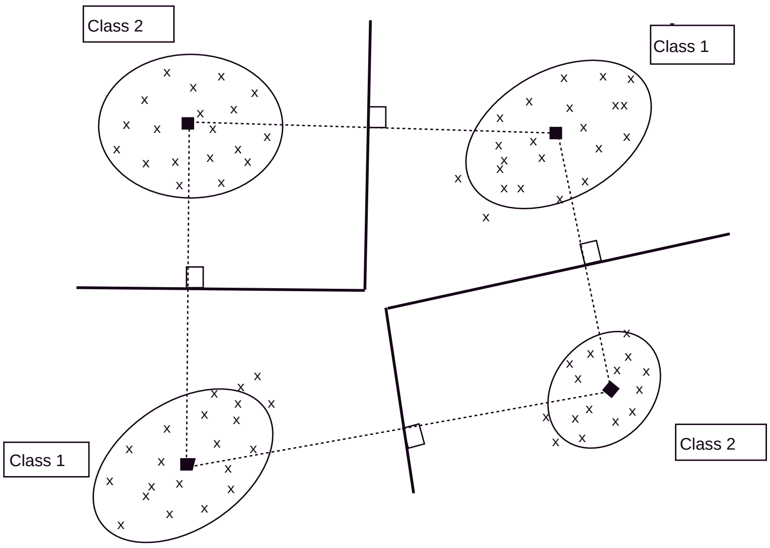
Bemærk at sidste led kan udregnes på forhånd for alle prototyper, og ikke behøver at blive evalueret ved klassifikation af hver ny observation $\mathbf{x}$. Hvis klasserne ikke er lineært separable, men visse klasser kan opdeles i underklasser således at problemet herved bliver lineært separabelt kan diskriminantfunktionen formuleres ved:

$$d_i(\mathbf{x}) \;=\; -\; \min_{j \in [1..N_i]} dist(\mathbf{x}, \mathbf{x}_{ij}) \tag{5.4}$$

hvor N_i er antallet af underklasser af klasse i. Benyttes det euclidiske afstandsmål bliver diskriminantfunktionen:

$$d_i(\mathbf{x}) = \max_{j \in [1..N_i]} \left\{ (\mathbf{x}^t \mathbf{x}_{ij} - \frac{1}{2}(\mathbf{x}_{ij})^t)\mathbf{x}_{ij} \right\} \tag{5.5}$$

Som illusteret nedenstående vil denne formulering give anledning til et sæt af stykvis lineære diskriminantfunktioner.



Figur 5.5: Stykvis lineære diskriminantfunktioner for et 2-klasse problem hvor hver klasse er beskrevet ved 2 prototyper.

5.3 Statistisk klassifikation

I mange situationer varierer de ekstraherede features meget i forhold til afstandene mellem prototyperne. I disse tilfælde er det ofte umuligt at bestemme grænsefunktioner der separere klasserne fuldstændigt. Den deterministiske tilgang, beskrevet i forrige afsnit, søger at opnå en klassifikationsfejl på 0 (nul), og har kun begrænset værdi i disse tilfælde. I stedet er det nødvendigt at bestemme klassifikationsregler, der er statistisk bestemt. I dette afsnit beskrives en statistisk optimal klassifikation. Denne har meget bred anvendelse, og betragtes som en reference, i forhold til hvilken nye metoder kan sammenlignes.

Beslutningsprocessen i et system til mønstergenkendelse kan betragtes som et statistisk spil , hvor systemet spiller mod "naturen". Denne antages at kende klasserne hvorfra data kan stamme. Hver runde af spillet foregår ved at naturen vælger klasse ω_i, samt en observation $\mathbf{x}$ fra denne klasse. Systemet skal, præsenteret for $\mathbf{x}$ "gætte" klassen som observationen stammer fra. Hvert korrekt gæt belønnes med en vis størrelse, og hvert forkert gæt straffes med en anden størrelse. Spillet antages at være en såkaldt "0-sum spil", hvor den ene spillers tab er lig den anden spillers gevinst. Opgaven for systemet vil typisk være at finde en (optimal) beslutningsproces, der minimerer det gennemsnitlige tab. En alternativ strategi, som vi ikke skal omtale yderligere, er at minimere det maksimale tab.

Antag at systemet klassificerer $\mathbf{x} \in \omega_j$ når $\mathbf{x}$ faktisk stammede fra ω_i. *Genvinst-* eller *tabs-funktionen* L_{ij} angiver da udgiften (eller gevinsten) for systemet. Som konvension vil vi betegne tab med positive værdier, og gevinster med ikke-positive værdier. Antag at antallet af klasser er M. L_{ij} kan da opskrives som en matrice med dimension $M \times M$, hvis antallet af faktiske klasser svarer til det antal, som systemet opererer med. Som nævnt vælger "naturen" først en klasse ω_i. Dette valg foretages ud fra den a priori sandsynlighed $p(\omega_i)$ for klassen, dvs. ud fra hyppigheden af forekomster af elementer fra denne klasse. Derefter vælges en observation $\mathbf{x}$ fra ω_i. Det antages at "naturen" ikke er en intelligent modspiller (der forsøger at maksimere egen gevinst), men vælger blindt på basis af de a priori sandsynligheder. Derimod tillades at systemet udspionere naturen, ved at konstruere eksperimenter, der skal kortlægge antallet af klasser, samt variabiliteten af $\mathbf{x}$ for hver klasse. Denne proces, der tænkes at gå forud for spillet, har vi diskuteret i afsnittet om klyngeanalyse.

Antag at sandsynligheden for at $\mathbf{x}$ stammer fra klasse ω_i er bestemt ved $p(\omega_i|\mathbf{x})$. Det forventede tab (for systemet) ved klassifikation af $\mathbf{x}$ til ω_j er da givet ved:

$$r_j(\mathbf{x}) \;=\; \sum_{i=1}^{M} L_{ij}\; p(\omega_i|\mathbf{x}) \tag{5.6}$$

Udtrykket kaldes også *det betingede gennemsnitlige tab.* Systemet kan vælge at klassificere $\mathbf{x}$ til enhver af de M klasser. Metoden er derfor at beregne det gennemsnitlige tab $r_1(\mathbf{x}), r_2(\mathbf{x}), \ldots, r_M(\mathbf{x})$ for alle M muligheder, og at vælge den klassifikation der resulterer i det mindste tab. Dette svarer til klassifikationsfunktionen:

$$c(\mathbf{x}) \;=\; j \quad \text{hvis} \quad r_j(\mathbf{x}) < r_i(\mathbf{x}) \quad \forall i \neq j, i \in [1..M] \tag{5.7}$$

Den klassifikator, der minimerer det totale forventede tab kaldes en *Bayes klassifikator* . Denne er, statistisk set, derfor optimal. Fra Bayes regel :

$$p(\omega_i|\mathbf{x}) \;=\; \frac{p(\mathbf{x}|\omega_i)p(\omega_i)}{p(\mathbf{x})}$$

har vi at:

$$r_j(\mathbf{x}) \;=\; \sum_{i=1}^{M} L_{ij}\; p(\mathbf{x}|\omega_i)p(\omega_i) \tag{5.8}$$

idet konstanten $p(\mathbf{x})$ er en fælles faktor for alle funktionerne $r_j(\mathbf{x})$, og derfor ingen indflydelse har på valget af klassifikationen. Funktionen $p(\mathbf{x}|\omega_i)$ kaldes for *likelihood funktionen* for klasse ω_i.

Lad os betragte situationen $M = 2$. Vi har da:

$$\begin{aligned} r_1(\mathbf{x}) &= L_{11}p(\mathbf{x}|\omega_1)p(\omega_1) + L_{21}p(\mathbf{x}|\omega_2)p(\omega_2) \\ r_2(\mathbf{x}) &= L_{12}p(\mathbf{x}|\omega_1)p(\omega_1) + L_{22}p(\mathbf{x}|\omega_2)p(\omega_2) \end{aligned}$$

Observationen $\mathbf{x}$ tilordnes klassen ω_1 hvis $r_1(\mathbf{x}) < r_2(\mathbf{x})$, dvs. hvis:

$$L_{11}p(\mathbf{x}|\omega_1)p(\omega_1) + L_{21}p(\mathbf{x}|\omega_2)p(\omega_2) \quad < \quad L_{12}p(\mathbf{x}|\omega_1)p(\omega_1) + L_{22}p(\mathbf{x}|\omega_2)p(\omega_2)$$

eller, ækvivalent, hvis:

$$(L_{21} - L_{22})p(\mathbf{x}|\omega_2)p(\omega_2) \quad < \quad (L_{12} - L_{11})p(\mathbf{x}|\omega_1)p(\omega_1) \tag{5.9}$$

Vi kan generelt antage at $L_{ij} > L_{ii}$, fordi tabet ved en fejlklassifikation må være større end ved en korrekt klassifikation. Alle størrelserne i ovenstående udtryk er derfor positive. Vi kan videre skrive betingelsen:

$$l_{12} \quad = \quad \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} \quad > \quad \frac{p(\omega_2)}{p(\omega_1)}\,\frac{L_{21} - L_{22}}{L_{12} - L_{11}} \quad = \quad \theta_{12} \tag{5.10}$$

Venstresiden af dette udtryk, dvs. forholdet mellem de to likelihood funktioner, betegnes *likelihood ratioen* og betegnes l_{12}. Højresiden af udtrykket kaldes *grænseværdien* (eng. threshold value) og betregnes θ_{12}. Bayes beslutningsregel for $M = 2$ bliver derfor:

1. Vælg klasse ω_1 hvis $l_{12} > \theta_{12}$.
2. Vælg klasse ω_2 hvis $l_{12} < \theta_{12}$.
3. Vælg klasse tilfældigt hvis $l_{12} = \theta_{12}$.

I det generelle tilfælde ($M > 2$) defineres tilsvarende:

$$l_{ij} \quad = \quad \frac{p(\mathbf{x}|\omega_i)}{p(\mathbf{x}|\omega_j)} \tag{5.11}$$

$$\theta_{ij} \quad = \quad \frac{p(\omega_j)}{p(\omega_i)}\,\frac{L_{ji} - L_{jj}}{L_{ij} - L_{ii}} \tag{5.12}$$

Klassifikationsfunktionen bliver

$$c(\mathbf{x}) \;=\; i \quad \text{hvis} \quad l_{ij} > \theta_{ij} \quad \forall j \neq i, \;\; j \in [1..M] \tag{5.13}$$

Bayes klassifikator, som beskrevet ovenfor, er nok den mest benyttede klassifikator i mønstergenkendelse. Der findes imidlertid en række alternative metoder, der er baseret på andre statisktiske beslutningskriterier. Ved *minimax* kriteriet ønskes det gennemsnitlige tab under de værst tænkelige omstændigheder minimeret. Et andet kriterie kaldes *Neyman-Pearson* kriteriet. Vi skal ikke her diskutere disse kriterier nærmere (se evt. [5]), blot notere, at for begge alterative metoder er likelihood ratioen l_{ij} den samme, hvorimod grænseværdien θ_{ij} er forskellig.

Eksempel
Antag at et binært signal sendes over en støjfyldt kanal. De enkelte dataelementer er altså enten 0 eller 1. Antag videre at den støj der tilføres signalet er additiv og kan modelleres ved en normalfordeling med middelværdi 0 og spredning σ. Lad ω_1 være hypotesen at et 0 var sendt, og ω_2 være hypotesen at et 1 var sendt. Lad yderligere $p(0)$ hhv. $p(1)$ være de priori sandsynligheder for afsendelse af et 0 hhv. et 1. Vi har da at:

$$\begin{aligned} p(x|\omega_1) &= \frac{1}{\sqrt{2\pi}\sigma}\ exp(-\frac{x^2}{2\sigma^2}) \\ p(x|\omega_2) &= \frac{1}{\sqrt{2\pi}\sigma}\ exp(-\frac{(x-1)^2}{2\sigma^2}) \end{aligned}$$

Likelihood ratioen bliver derfor:

$$l_{12}(x) \ = \ exp(\frac{1-2x}{2\sigma^2})$$

Ved Bayes klassifikation vil x blive fortolket som et 0 hvis $l_{12} > \theta_{12}$, dvs. hvis:

$$x \ < \ \frac{1}{2} - \sigma^2 ln(\theta_{12}) \ = \ \frac{1}{2} - \sigma^2 ln(\frac{L_{21}p(1)}{L_{12}p(0)})$$

Hvis kanalen er støjfri, dvs. hvis $\sigma = 0$, vil x blive fortolket som et 0 hvis $x < 0.5$. Dette svarer til intuitionen. Tilsvarende gælder hvis $\theta_{12} = 1$. I andre tilfælde er den optimale grænseværdi ikke i overensstemmelse med intuitionen.

Eksempel slut

I mange situationer er bestemmelse af tabs-funktionen essentiel for den optimale løsning. Ved klassifikation af flyvemaskiner som fjendtlige eller venlige kan klassifikation af en fjendtlig flyver som venlig være katastrofal, hvorimod klassifikation af en venlig flyver som fjende kan have mindre dramatiske konsekvenser. Et andet eksempel er kvalitetskontrol i en industriel proces. Her kan det koste dyrt at klassificere et dårligt produkt som godt, medens den omvendte fejlklassifikation sjældent er vigtig. Den sidste type af fejl kan ofte korrigeres ved efterfølgende manuelt check af de frasorterede produkter. Konsekvenserne af de forskellige typer af fejl er således generelt forskellige. I resten af dete kapitel skal vi imidlertid antage at alle fejl er lige slemme. Denne antagelse kan modelleres ved:

$$L_{ij} \ = \ 1 \ - \ \delta_{ij} \ = \begin{cases} 1 & \text{hvis} \quad i \neq j \\ 0 & \text{hvis} \quad i = j \end{cases} \tag{5.14}$$

Ved denne definition er tabet for en korrekt klassifikation lig 0. Tabet for enhver fejlagtig klassifikation er 1. Dette betyder at *fejlraten* (det betingede gennemsnitlige tab) bliver:

$$\begin{aligned} r_j(\mathbf{x}) &= \sum_{i=1}^{M}(1-\delta_{ij})\ p(\mathbf{x}|\omega_i)p(\omega_i) && (5.15) \\ &= p(\mathbf{x}) \ - \ p(\mathbf{x}|\omega_j)p(\omega_j) && (5.16) \end{aligned}$$

Bayes klassifikationen, der i dette tilfælde vil minimere sandsynligheden for fejl i klassifikationen, bliver derfor:

$$c(\mathbf{x}) = i \quad \text{hvis} \quad p(\mathbf{x}|\omega_i)p(\omega_i) \; > \; p(\mathbf{x}|\omega_j)p(\omega_j) \quad \forall j \neq i, \;\; 1 \leq j \leq M \tag{5.17}$$

Som det huskes fra afsnittet om diskriminantfunktioner udtrykker ovestående regel blot at disse er bestemt ved:

$$d_i(\mathbf{x}) \; = \; p(\mathbf{x}|\omega_i)p(\omega_i) \tag{5.18}$$

I ovenstående udtryk for diskriminantfunktionerne kræves der eksplicit kendskab til både den betingede sandsynlighed $p(\mathbf{x}|\omega_i)$ og til den a priori sandsynlighed $p(\omega_i)$. I udledningen udnyttede vi Bayes regel. Tager vi alternativt udgangspunkt i (5.6) direkte, fremkommer en helt ækvivalent formulering:

$$d_i(\mathbf{x}) \; = \; p(\omega_i|\mathbf{x}) \tag{5.19}$$

Valget mellem hvilken af de to formuleringer (5.18) eller (5.19), der er mest hensigtsmæssig, afhænger af hvilke fordelinger der, i en given anvendelse, er kendte. I mange situationer foretages en indledende dataanalyse (herunder klyngeanalyse) på et sæt af træningsdata. Herved kan de a priori sandsynligheder estimeres ved frekvensen af observationer inden for hver klasse. Yderligere er det ofte muligt at modellere de betingede sandsynligheder $p(\mathbf{x}|\omega_i)$ ved en kendt parametrisk fordelingsfunktion (feks. en normalfordeling). Vi skal diskutere dette tilfælde nærmere i nedenstående afsnit. Sandsynlighederne $p(\omega_i|\mathbf{x})$ er umiddelbart sværere at estimere, men kan være fordelagtige ved brug i metoder hvor diskriminantfunktionerne indlæres inkrementalt. Vi skal ikke her diskutere dette nærmere, men referere til f.eks. [14,3] for nærmere diskussion.

5.3.1 Diskriminantanalyse for normalt fordelte mønstre

I dette afsnit beskrives Bayes klassifikatoren, bestemt ved diskriminantfunktionen (5.18) for normale fordelingsfunktioner $p(\mathbf{x}|\omega_i)$. Antagelsen om normalitet af fordelingen $p(\mathbf{x}|\omega_i)$ er dels begrundet i analytisk medgørlighed, dels fordi denne fordeling er en god approksimation til mange faktisk forekomne empiriske fordelinger. Fordelingen er givet ved:

$$p(\mathbf{x}|\omega_i) \; = \; \frac{1}{(2\pi)^{n/2}|\mathbf{C}_i|^{1/2}} \quad exp[-\frac{1}{2}(\mathbf{x}-\mathbf{m}_i)^t\mathbf{C}_i^{-1}(\mathbf{x}-\mathbf{m}_i)] \tag{5.20}$$

hvor $\mathbf{m}_i$ er middelværdivektoren for observationerne hørende til klasse i og $\mathbf{C}_i$ er den tilsvarende kovariansmatrice. Diskriminantfunktionen er givet ved: $d_i(\mathbf{x}) = p(\mathbf{x}|\omega_i)p(\omega_i)$. På grund af den eksponentielle form af normalfordelingen er det hensigstsmæssigt at arbejde med logaritmen af denne. Logaritmefunktionen er monotont voksende og ænder derfor ikke på sammenligningerne af diskriminantfunktionerne. Vi får således:

$$\begin{aligned} d_i(\mathbf{x}) \; &= \; ln \; [p(\mathbf{x}|\omega_i)p(\omega_i)] \qquad (5.21) \\ &= \; ln \; p(\mathbf{x}|\omega_i) \; + \; ln \; p(\omega_i) \\ &= \; ln \; p(\omega_i) \; - \; \frac{n}{2} ln \; 2\pi \; - \; \frac{1}{2} ln \; |\mathbf{C}_i| \; - \; \frac{1}{2}[\mathbf{x}-\mathbf{m}_i]^t\mathbf{C}_i^{-1}[\mathbf{x}-\mathbf{m}_i] \end{aligned}$$

I dette udtryk kan vi ignorere konstanten $\frac{n}{2} ln 2\pi$, fordi dette udtryk er uafhængigt af i. Diskriminantfunktionerne bliver herved:

$$d_i(\mathbf{x}) \;=\; ln\; p(\omega_i) \;-\; \frac{1}{2} ln|\mathbf{C}_i| \;-\; \frac{1}{2}[\mathbf{x}-\mathbf{m}_i]^t\mathbf{C}_i^{-1}[\mathbf{x}-\mathbf{m}_i] \qquad i=1,\ldots,M \qquad (5.22)$$

Ovenstående udtryk angiver Bayes diskriminantfunktionerne for normalt fordelte mønstre. Væsentlig ved udledningen af udtrykket var, ud over antagelsen om normalitet, antagelsen at tabs-funktionen var givet ved (5.14). Det ses af udtrykket, at diskriminantfunktionerne er anden ordens polynomier i $\mathbf{x}$. Det bedste en Bayes klassifikator kan gøre er således at skille klasserne ved denne type funktioner. Dette er imidlertid den optimale form, hvis klasserne virkelig er normalfordelte. Det er desuden det teoretisk optimale valg for en lang række af andre fordelingsfunktioner (se f.eks. [14]). Nedenstående skitseres en række specialtilfælde:

Antag først at kovariansmatricerne er ens for alle klasser, dvs. $\mathbf{C}_i = \mathbf{C}_j = \mathbf{C}$. Ved at ignorere alle led, der er uafhængige af i, fra diskriminantfunktionen fås:

$$d_i(\mathbf{x}) \;=\; lnp(\omega_i) \;+\; \mathbf{x}^t\mathbf{C}^{-1}\mathbf{m}_i \;-\; \frac{1}{2}\,\mathbf{m}_i^t\mathbf{C}^{-1}\mathbf{m}_i \qquad i=1,\ldots,M \qquad (5.23)$$

Det ses at diskriminantfunktionerne i dette tilfælde bliver lineære. Hvis der yderligere gælder at kovariansmatricen er lig identiteten, $\mathbf{C} = \mathbf{I}$, samt at de a priori sandsynligheder for at observationerne stammer fra de M klasser er ens, dvs. $p(\omega_i) = 1/M$, da reduceres udtrykket yderligere til:

$$d_i(\mathbf{x}) \;=\; \mathbf{x}^t\mathbf{m}_i \;-\; \frac{1}{2}\,\mathbf{m}_i^t\mathbf{m}_i \qquad i=1,\ldots,M \qquad (5.24)$$

Dette udtryk svarer til diskriminantfunktionerne for minimal afstands klassifikation, under brug af det euclidiske afstandsmål, hvor hver prototype er bestemt ved middelværdivektoren $\mathbf{m}_i$.

Eksempel
Det er vanskeligt at give et overskueligt eksempel på ovenfor beskrevne statistiske tilgang til klassifikationsproblemet, fordi dette (for at være realistisk) kræver at antallet af stikprøver er stor. Ikke desto mindre gives her et sådant eksempel. Betragt nedenstående grupper af 12 observationer, hver beskrevet ved en featurevektor af dimension 3, og hver manuelt klassificeret til én af 2 klasser.

observation	klasse	x_1	x_2	x_3
1	ω_1	3	0	0
2	ω_1	4	3	1
3	ω_1	5	1	2
4	ω_1	6	1	0
5	ω_1	7	0	0
6	ω_1	5	1	3
7	ω_2	-1	15	0
8	ω_2	0	9	1
9	ω_2	1	3	3
10	ω_2	1	10	5
11	ω_2	-1	1	7
12	ω_2	0	4	8

På basis af ovenstående observationer antages, at de a priori sandsynligheder for de to klasser er ens, dvs. $p(\omega_1) = p(\omega_2) = \frac{1}{2}$. Ved hjælp af de i kapitel 2 beskrevne metoder kan vi estimere middelværdivektorene og kovariansmatricerne til:

$$\mathbf{m_1} = (5,1,1) \qquad \text{og} \qquad \mathbf{m_2} = (0,7,4)$$

$$\mathbf{C}_1 = \frac{1}{5}\begin{pmatrix} 160 & 28 & 29 \\ 28 & 12 & 8 \\ 29 & 8 & 14 \end{pmatrix} \quad \text{og} \quad \mathbf{C}_2 = \frac{1}{5}\begin{pmatrix} 4 & -3 & 1 \\ -3 & 432 & 107 \\ 1 & 107 & 148 \end{pmatrix}$$

Disse matricer har de inverse:

$$\mathbf{C}_1^{-1} \approx \begin{pmatrix} 0.0607 & -0.0934 & -0.0724 \\ -0.0934 & 0.8168 & -0.2732 \\ -0.0724 & -0.2732 & 0.6632 \end{pmatrix} \quad \text{og} \quad \mathbf{C}_2^{-1} \approx \begin{pmatrix} 1.2645 & 0.0133 & -0.0181 \\ 0.0133 & 0.0142 & -0.0104 \\ -0.0181 & -0.0104 & 0.0414 \end{pmatrix}$$

og der gælder at $|\mathbf{C}_1| = 1712.8$ og $|\mathbf{C}_2| = 41508.4$. Vi har derfor at:

$$ln\ p(\omega_1) - \frac{1}{2}ln|\mathbf{C}_1| \approx -4.4161$$
$$ln\ p(\omega_2) - \frac{1}{2}ln|\mathbf{C}_2| \approx -6.0100$$

Lad $\mathbf{x} = (x_1, x_2, x_3)$ være en ny observation. Bayes diskriminantfunktioner er da givet ved:

$$d_1(\mathbf{x}) = -4.4161 - 0.5 \begin{bmatrix} x_1 - 5.0 \\ x_2 - 1.0 \\ x_3 - 1.0 \end{bmatrix}^t \begin{pmatrix} 0.0607 & -0.0934 & -0.0724 \\ -0.0934 & 0.8168 & -0.2732 \\ -0.0724 & -0.2732 & 0.6632 \end{pmatrix} \begin{bmatrix} x_1 - 5.0 \\ x_2 - 1.0 \\ x_3 - 1.0 \end{bmatrix}$$

og

$$d_2(\mathbf{x}) = -6.0100 - 0.5 \begin{bmatrix} x_1 - 0.0 \\ x_2 - 7.0 \\ x_3 - 4.0 \end{bmatrix}^t \begin{pmatrix} 1.2645 & 0.0133 & -0.0181 \\ 0.0133 & 0.0142 & -0.0104 \\ -0.0181 & -0.0104 & 0.0414 \end{pmatrix} \begin{bmatrix} x_1 - 0.0 \\ x_2 - 7.0 \\ x_3 - 4.0 \end{bmatrix}$$

Antag nu konkret at $\mathbf{x} = (x_1, x_2, x_3) = (2, 5, 4)$. Ved indsættelse og udregning fås at:

$$\begin{aligned} d_1(\mathbf{x}) &= -12.05 \\ d_2(\mathbf{x}) &= -8.54 \end{aligned}$$

Bayes klassifikationsfunktion vil derfor anbefale, at klassificere den ny observation til klasse ω_2, da $d_2(\mathbf{x}) > d_1(\mathbf{x})$.

Eksempel slut

Eksempel
For at kunne foretage automatisk sortering af en række produkter, her søm, skruer, og møtrikker, har en fabrik overvejet et system, hvor et kamera overvåger et transportbånd, hvor produkterne bringes forbi et for et. Hvert billede analyseres således at produktet ekstraheres fra baggrunden. Derpå bestemmes tre mål, der karakteriserer produktets silhouet. De tre mål er længde, kompakthed, og regularitet. På et sæt af træningsdata (hvor den korrekte klassifikation er kendt, er følgende middelværdier for de tre features bestemt:

Middeværdier $\mathbf{m}_i$

feature	søm	skruer	møtrikker
længde	32.4	29.1	9.6
kompakthed	10.8	8.4	1.3
regularitet	5.2	17.1	6.0

I det følgende refererer indeks 1 til klassen søm, indeks 2 til klassen skruer, og indeks 3 til klassen møtrikker. I estimationen af middelværdierne benyttede man billeder af $n_1 = 27$ søm, $n_2 = 34$ skruer, og $n_3 = 20$ møtrikker. Det antages at kovariansmatricerne for de tre objekter er ens, dvs. $\mathbf{C}_i = \mathbf{C}$ for $i = 1, 2, 3$. Denne er estimeret ved:

$$\mathbf{C} = \frac{1}{n_1 + n_2 + n_3 - 3} \left[\sum_{i=1}^{n_1} (\mathbf{x}_1^i - \mathbf{m}_1)(\mathbf{x}_1^i - \mathbf{m}_1)^t + \sum_{i=1}^{n_2} (\mathbf{x}_2^i - \mathbf{m}_2)(\mathbf{x}_2^i - \mathbf{m}_2)^t + \sum_{i=1}^{n_3} (\mathbf{x}_3^i - \mathbf{m}_3)(\mathbf{x}_3^i - \mathbf{m}_3)^t \right]$$

hvor $\mathbf{x}_j^i$ er den i'te observation (søjle featurevektor) i den j'te klasse. Ved estimationen blev $\mathbf{C}$ bestemt til:

Kovariansmatrice C

feature	længde	kompakthed	regularitet
længde	9.0	12.1	2.7
kompakthed	12.1	7.2	9.4
regularitet	2.7	9.4	3.5

Vi er nu interesseret i at få opstillet diskriminantfunktioner således at fremtidige observationer kan klassificeres til en af de 3 klasser. Først vil det dog være rimeligt at undersøge om de tre egenskaber (features) overhovedet er forskellige for de tre populationer. Vi må altså undersøge om $\mathbf{m}_i = \mathbf{m}_j$ for $i \neq j$. Mahalanobis afstanden mellem de tre prototyper er givet ved:

$$D_{i,j}^2 = (\mathbf{m}_i - \mathbf{m}_j)^t \, \mathbf{C}^{-1} \, (\mathbf{m}_i - \mathbf{m}_j)$$

Ved udregning fås:

Mahalanobis afstande

$D^2_{1,2}$	$D^2_{1,3}$	$D^2_{2,3}$
32.39	78.40	9.46

Statistikeren Fischer, der parallelt med Mahalanobis, udviklede diskriminantanalysen, opstillede en teststørrelse Z for hypotesen at to middelværdier var ens. Z er givet ved:

$$Z_{ij} \;=\; \frac{n_i + n_j - p - 1}{p(n_i + n_j - 2)} \cdot \frac{n_i n_j}{n_i + n_j} D^2_{ij}$$

Bemærk, at dette udtryk primært adskiller sig fra udtrykket i ligning (1.18) ved brug af mahalanobisafstanden. I ligning (1.18) benyttes en diagonalmatrice (med skalerede varianser) i stedet for kovariansmatricen. Under hypotesen H_0 at $\mathbf{m}_i = \mathbf{m}_j$ er Z fordelt som F(p, $n_i + n_j - p - 1$). Hvis Z er større end den ved F-fordelingen forudsagte (på det valgte konfidensniveau) kan hypotesen H_0 afvises. Idet $p = 3$ og vi benytter 0.99 konfidensniveauet fås:

Testværdier

(i,j)	Z_{ij}	frihedsgrader	F
(1,2)	156.96	(3,59)	4.1
(1,3)	286.91	(3,43)	4.3
(2,3)	38.19	(3,50)	4.2

Vi kan heraf se, at alle tre hypoteser kan forkastes. Det giver derfor god mening, at forsøge at opstille de tre diskriminatorer. Da der ikke foreligger oplysninger om de a priori sandsynligheder for forekomster af observationer fra de tre klasser antages disse at være ens. Dette leder til diskriminantfunktionerne:

$$d_i(\mathbf{x}) \;=\; \mathbf{x}^t\mathbf{C}^{-1}\mathbf{m}_i \;-\; \frac{1}{2}\,\mathbf{m}_i^t\mathbf{C}^{-1}\mathbf{m}_i$$

Som beskrevet tidligere ses at diskriminantfunktionerne er lineære. Ved indsættelse fås:

$$\begin{aligned}
d_1 &= 3.35\,x_1 + 1.08\,x_2 - 3.99\,x_3 - 49.71 \\
d_2 &= 0.71\,x_1 + 2.27\,x_2 - 1.76\,x_3 - 4.82 \\
d_3 &= 0.12\,x_1 + 0.85\,x_2 - 0.67\,x_3 + 0.88
\end{aligned}$$

hvor $\mathbf{x} = (x_1, x_2, x_3)$ angiver featurevektoren. Har vi eksempelvis foretaget en måling $(14.7, 5.2, 7.0)$ fås diskriminantværdierne $\mathbf{d} = (-22.78, 5.10, 2.39)$. Vi vil derfor klassifcere observationen som værende en skrue.

Eksempel slut

5.4 Opgaver

1. Vis at ingen observationer i det grå område i figur 5.2 kan klassificeres ved brug af de i figuren viste beslutningsfunktioner.

2. Lad A = (5,7), B = (-2,5), og C = (1,-1) være tre prototyper af klasser for mønstre karakteriseret ved to features. Opstil diskriminantfunktionerne ved brug af mindste afstands klassifikatorer og det euclidiske afstandsmål. Tegn funktionerne.

3. Med udgangspunkt i eksemplet om transmition af signaler skal likelihoodfunktionen l_{12} udregnes under antagelse at støjen er eksponentialfordelt med parameter λ. Tegn funktionen og sammenlig med den i eksemplet udledte funktion. Antag at $\theta_{12} = 1$. Hvad bliver grænsen for klassifikation mellem de to prototyper.

4. Antag at der er givet 3 prototyper (i 2 features), hver karakteriseret ved deres middeværdivektor of deres estimerede kovariansmatrice. Lad disse være givet ved:

$$\mathbf{m}_1 = (6,7) \quad \mathbf{m}_2 = (3,5) \quad \mathbf{m}_3 = (7,3)$$

$$\mathbf{C}_1 = \begin{pmatrix} 0.6 & 0.0 \\ 0.0 & 0.6 \end{pmatrix} \quad \mathbf{C}_2 = \begin{pmatrix} 1.0 & -1.0 \\ -1.0 & 1.6 \end{pmatrix} \quad \mathbf{C}_3 = \begin{pmatrix} 0.8 & 0.5 \\ 0.5 & 1.0 \end{pmatrix}$$

   Antag at alle tre mønstre er normalfordelte. Opstil diskriminantfunktionerne for Bayes klassifikationen. Afgør hvilke klasser som stikprøverne (3,2), (6,5), og (10,6) vil blive klassificeret til.

5. Antag at der er givet 3 prototyper (i 2 features), hver karakteriseret ved deres middeværdivektor of deres estimerede kovariansmatrice. Lad disse være givet ved:

$$\mathbf{m}_1 = (3,2) \quad \mathbf{m}_2 = (9,2) \quad \mathbf{m}_3 = (6,7)$$

$$\mathbf{C}_1 = \begin{pmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{pmatrix} \quad \mathbf{C}_2 = \begin{pmatrix} 1.0 & 0.0 \\ 0.0 & 1.0 \end{pmatrix} \quad \mathbf{C}_3 = \begin{pmatrix} 2.0 & 0.0 \\ 0.0 & 1.0 \end{pmatrix}$$

   Antag at alle tre mønstre er normalfordelte. Hvilken klasse vil stikprøven (6,3) blive klassificeret til:
   1) Klasse 1
   2) Klasse 2
   3) Klasse 3
   4) Ingen af de tre klasser
   5) Alle tre klasser
   6) En tilfældig af klasserne
   7) Det ved jeg ikke

Kapitel 6

Neurale Net

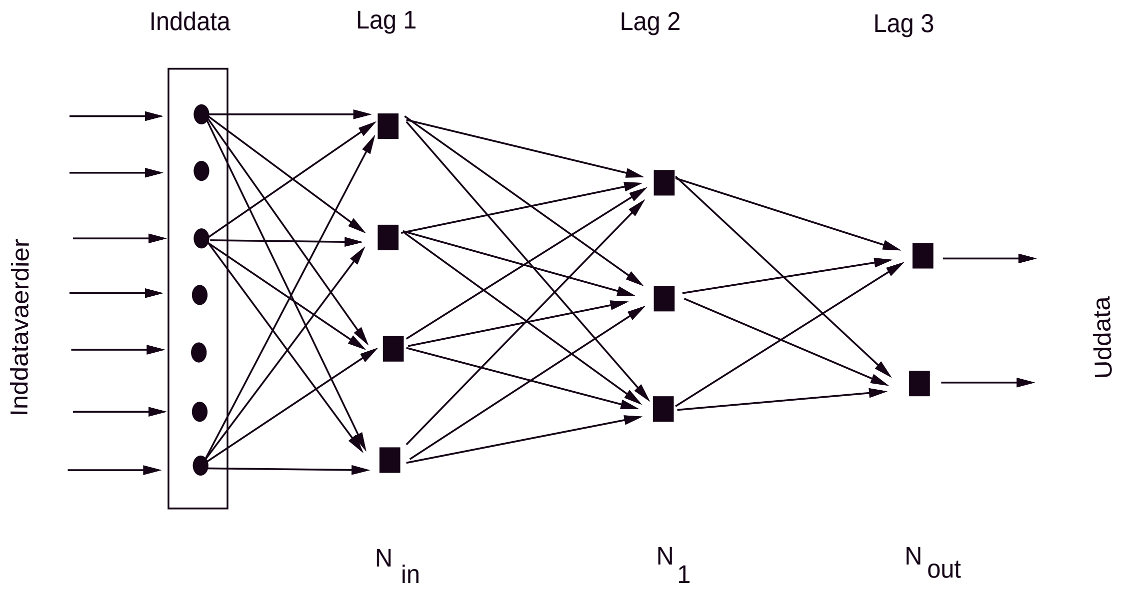
Brug af (kunstige) neurale net til klassifikation har været kendt de sidste 30-40 år, men har især været populære inden for de sidste 10 år. En grund til populariteten af neurale net er at de (i princippet) er lette at forstå, samt at de er meget generelle. I dette kapitel skal vi udelukkende diskutere anvendelsen af neurale net som klassifikatorer.

6.1 Lagdelte, fuldt forbundne net uden tilbagekobling

Et neuralt net er karakteriseret ved dets netstruktur og den beregningsmåde som neuronerne fungerer efter. Alle neuroner benytter principielt samme beregningsmåde. Der findes flere klasser af neurale net, eksempelvis de lagdelte, de med tilbagekobling, de fuldt forbundne osv. I det følgende skal vi udelukkende beskrive lagdelte, fuldt forbundne netværk uden tilbagekobling. Denne type er især anvendelige ved klassifikation af data. Andre nettyper, f.eks. fuldt forbundne net med tilbagekobling, kan være effektive f.eks. til klyngeanalyse. Når der i det følgende benyttes termen *neuron* skal der forstås en beregningsenhed, der givet N inddataværdier kan producere en uddataværdi. Vi skal senere beskrive beregningsmåden for en neuron. Figur 6.1 illustrerer et lagdelt fuldt forbundet net uden tilbagekobling.

Nettet består at N_{in} indata-neuroner, N_{out} uddata neuroner, samt et antal $N_1, N_2, \ldots, N_k$ af neuroner i de k mellemlag. Data flyder i én retning gennem nettet. Dette har ingen hukommelse. Som beskrevet senere er der til hver kant i nettet knyttet en vægt. Denne vægt kan justeres under indlæring af nettet, men er derefter konstant.

Antallet af neuroner i det første lag er i visse tilfælde bestemt på forhånd ved dimensionen af observationsvektoren, men kan være både større eller mindre. Antallet af neuroner i uddatalaget er bestemt ved antallet af klasser. Antallet af lag er typisk 1, 2 eller 3. Net med 3 lag af neuroner er de mest generelle. Antallet af neuroner i hver lag må generelt vælget efter den opgave nettet skal løse. Vi skal senere vende tilbage til dette problem.



Figur 6.1: Netstrukturen i et lagdelt neuralt net uden tilbagekobling. Nettet har et skjult lag, et inddatalag, og et uddata lag.

6.2 Beregningsmåde

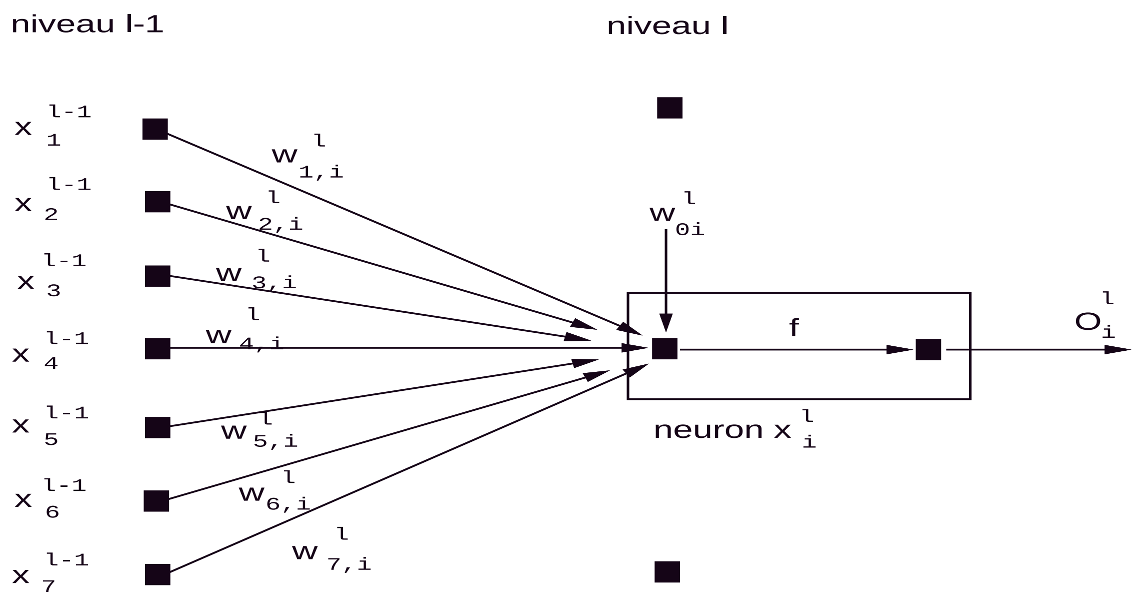
Stort set alle neurale net benytter samme beregningsmåde. Hver neuron beregner først en linearkombination af inddata og adderer en bias. På dette resultat beregnes uddataværdien fra neuronen dernæst ved anvendelse af en såkaldt grænsefunktion f. Lad x_i^l være den i'te neuron i det l'te lag i nettet, og lad w_{ik}^l være den vægt hvormed neuronen x_k^{l-1} påvirker linearkombinationsudregningen i neuron x_i^l. Da kan uddata O_i^l fra den i'te neuron i det l'te lag skrives:

$$O_i^l = f(w_{i0}^l + \sum_{k=1}^{N_{l-1}} w_{ik}^l \cdot x_k^{l-1}) \quad (6.1)$$

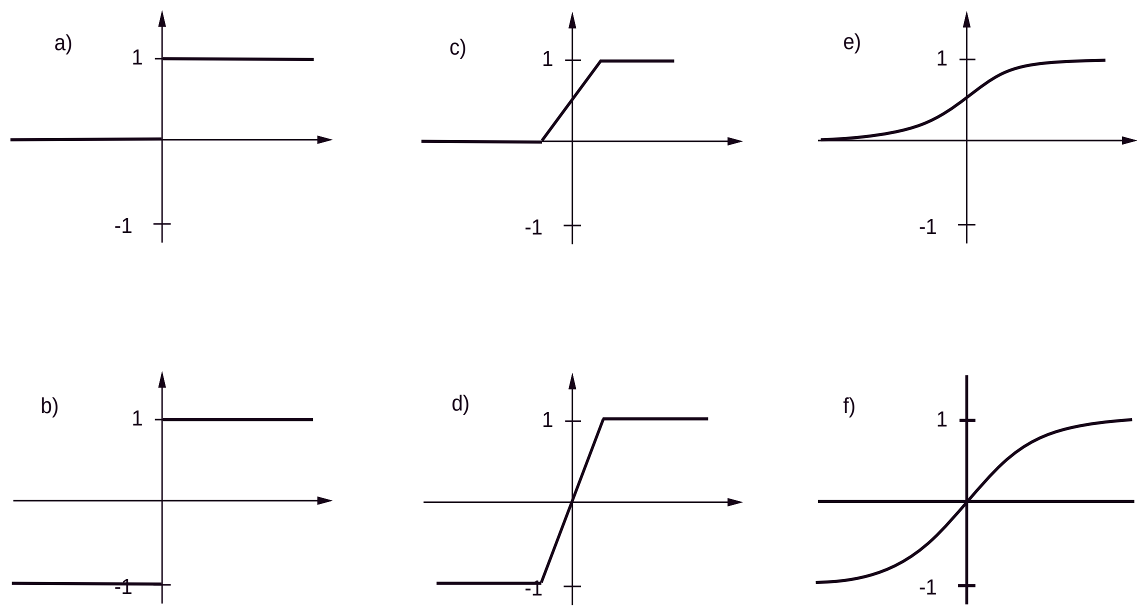
$$= f(\mathbf{w}_i^l \cdot \mathbf{x}^{l-1}) \quad (6.2)$$

hvor N_{l-1} er antal neuroner på det $(l-1)$'te lag, hvor w_{i0}^l er bias for neuronen x_i^l, hvor $\mathbf{x}^{l-1}$ er vektoren $(1, x_1^{l-1}, x_2^{l-1}, \ldots, x_{N_{j-1}}^{l-1})$, og hvor $\mathbf{w}_i^l = (w_{i0}^l, w_{i1}^l, \ldots, w_{iN_{j-1}}^l)$.

Funktionen f er typisk ulineær. Hvis f var lineær ville uddata fra hele nettet kunne beskrives ved én enkelt linearkombination af inddata. Denne situation er sjældent interessant. Nedenstående figur illustrerer nogle ofte benyttede grænsefunktioner f. Virkningen af grænsefunktionen er sammenligelig med en test. Hvis (værdien af) linearkombinationen er stor er uddata lig 1. Hvis (værdien af) linearkombinationen er stor negativ er uddata lig 0 eller -1 (afhængig af den valgte konvention). Hvis linearkombinationen er tæt på nul, er uddata bestemt ved formen af grænsefunktionen. Det kan her bemærkes at visse metoder til oplæring af neurale net forudsætter at grænsefunktionen er differentiabel. Væsenligt



Figur 6.2: Beregningsmåde for en neuron med $N_{l-1} = 7$.

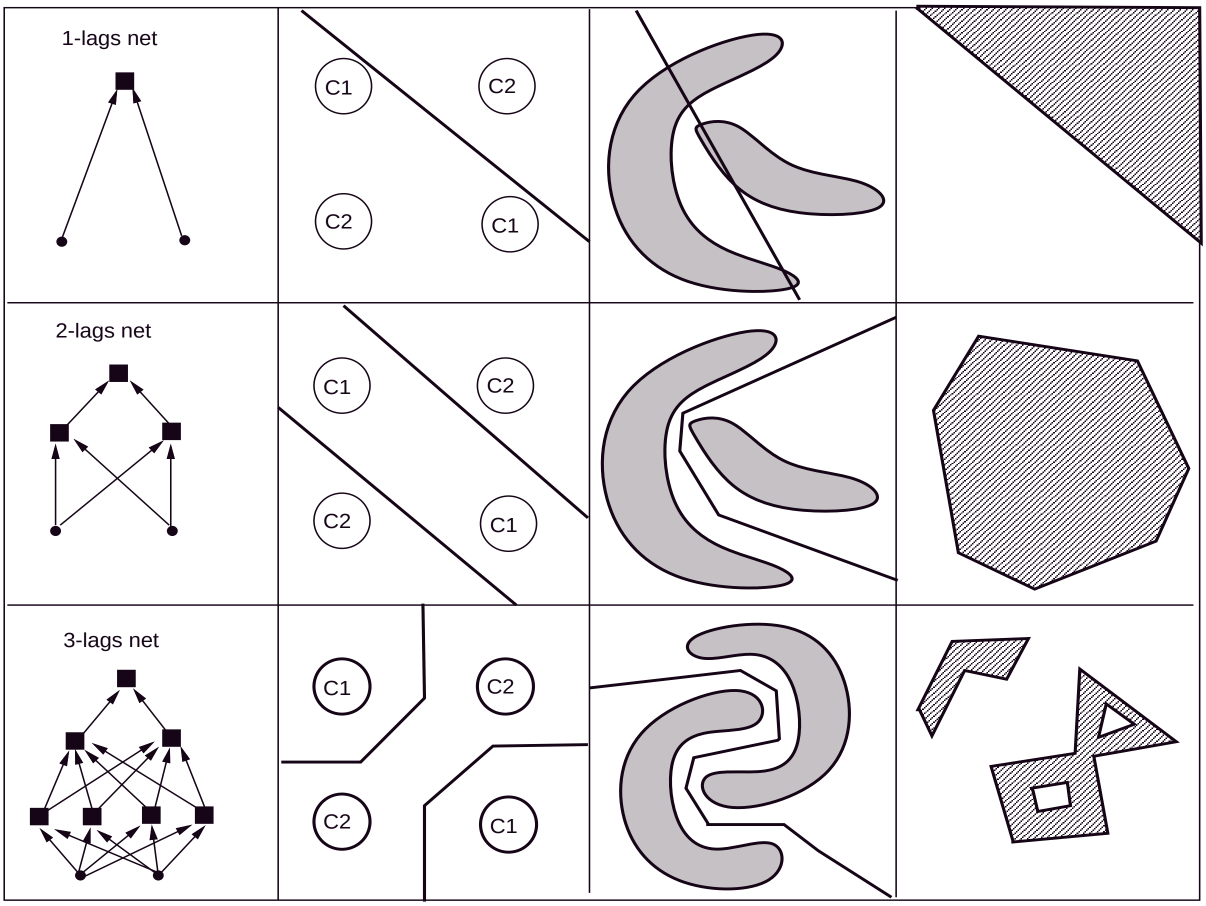


Figur 6.3: Ofte benyttede gænsefunktioner. Funktionerne i a) og b) er ikke kontinuerte, funktionerne i c) og d) er ikke differentiable. Funktionen i e) vælges typisk lig $\frac{1}{1+e^{-\alpha t}}$. Funktionen i f) kunne være $\tanh(\alpha t)$.

er, at den samme grænsefunktion anvendes i alle neuroner. Det er derfor vægtene $\mathbf{w}$, der bestemmer hvad nettet beregner.

6.3 Hvad kan et neuralt net modellere

Det er let at se, at ved valg af a) eller b) i figur 6.3 som grænsefunktion, da vil ovenfor beskrevne model af en neuron bestemme en linæær beslutningsfunktion. Hvis inddata skal klassificeres mht. en af to lineært seperable klasser vil dette således kunne gøres af blot én neuron. Anderledes sagt opdeles featurerummet i 2 halvrum. Hvis der benyttes et 2-lags neuralt net (dvs. et net uden mellemlag) vil hver neuron i inputlaget kunne definere et sådant halvrum. Output neuronen vil derefter kunne bestemme fællesmængde af et udvalg af disse halvrum. Heraf ses, at et 2-lags net vil kunne klassificere inddata hørende til konvekse delmængder af featurerummet. Hvis der benyttes et 3-lags net vil output neuronen kunne bestemme foreningsmængden af et antal konvekse delrum, beregnet af neuronerne i mellemlaget. Herved kan featurerummet opdeles i ikke-konvekse delrum. Dette er illustreret i nedenstående figur.



Figur 6.4: Klassifikation ved neurale net. Venstre søjle illustrerer netstrukturen. Anden søjle viser de beslutningsfunktioner som nettet kan implementere ved det såkaldte *exclusive or problem.* Tredie søjle og højre søjle illustrerer hvor komplekse opdelinger af featurerummet, som nettene kan foretage.

Det kan vises at evnen til at opdele featurerummet ikke bliver kvalitativt bedre ved at tilføje nye lag ud over 3 lag. Til klassifikationsformål benytter man derfor altid 3-lags net, medmindre et 2-lags net er tilstrækkeligt. Det ses også, at afgrænsningen af klasser altid vil ske ved stykvis lineære funktioner. Hvis formen af en klasse er krum vil der derfor skulle mange neuroner i mellemlagene til at modellere klassen godt. Dette er en væsenlig ulempe, fordi antallet af parametre (antal vægte) dermed stiger.

Det skal bemærkes, at der findes en lang række andre typer af neurale net, der er mere velegnede f.eks. til klyngeanalyse, til approksimation af funktionsværdier mv. Vi skal ikke beskrive disse nærmere, men dog nævne at man kan konstuere neurale net, der kan approksimere en funktion vilkårligt godt. Jo bedre en approksimation der kræves, jo flere neuroner er det nødvendigt at nettet indeholder. Neurale net er derfor meget generelle. Dette resultat svarer godt til intuitionen fra ovenstående figur. Vi kan approksimere formen af en klasse vilkårlig godt, blot vi har neuroner nok.

6.3.1 Antal parametre

Hvis antallet af neuroner i (det fuldt forbundne) net er stort, da vil antallet af parametre (vægte) også blive stort. Lad N_i være antal neuroner i det i'te lag, således at N_0 er antal input, og N_n er antal output neuroner. Da er det totale antal af parametre i systemet bestemt ved:

$$P = \sum_{k=1}^{n} N_k \cdot (N_{k-1} + 1) \tag{6.3}$$

Antag, som eksempel, at vi ønsker at genkende håndskrevne tal ud fra et 11×17 binært rasterbillede. Længden af inddatavektoren er derfor $11 \cdot 17 = 187$. Antallet af klasser er lig 10. Benyttes er 2-lags net med N_1 neuroner i det første lag (og 10 neuroner i det andet lag) vil der ialt være $198N_1 + 10$ vægte (parametre) i systemet. Det kan være meget svært at vælge et passende antal af neuroner. Dette afhænger bl.a. af kompleksiteten af det, som nettet skal beregne. En meget grov tommelfingerregel (der kun bør bruges hvis man ingen anden information har) er at bestemme denne som gennemsnittet af antal neuroner i input og output laget. Benyttes denne metode på antallet af inddataværdier i ovenstående eksempel bliver $N_1 = 99$, svarende til ialt 19.612 parametre i systemet. Dette er et meget stort tal. Selv om blot 10 neuroner er tilstrækkeligt i det første lag vil der ialt blive 1990 parametre. Benyttes et 3-lags net vil antallet af parametre blive langt større end de ovennævnte tal. Det er klart, at det i praksis vil være helt umuligt at angive alle disse parameterværdier korrekt på forhånd. Man er nødt til at estimere vægtene ved oplæring. Vi skal i et senere afsnit beskrive en simpel metode til indlæring af vægte.

6.3.2 Generaliseringsevne

Det er klart at jo flere vægte et net indeholder, jo mere kan nettet beregne. Ofte vil det imidlertid være en fordel at minimere antallet af vægte. Dette skyldes ikke kun problemet med estimation af vægtene, men er knyttet til nettets *generaliseringsevne*.

Problemet med generaliseringsevne er ikke kun relevant for neurale net. Antag at vi har givet n punkter (x_i, y_i), og at vi ønsker at fitte et polynomie $p(x)$ til disse data. Problemet er at vælge graden m af det approksimerende polynomium. Antag, som eksempel at m vælges lig $n-1 \gg 3$, og antag at de n punkter faktisk ligger pænt på en parabel. Da vil estimationen af parametrene i høj grad være bestemt af støjen i de data (Se også Kapitel 3). Ved brug af for mange parametre vil funktionen $p(x)$ ofte oscillere meget kraftigt. Mellem punkterne x_i vil $p(x)$ derfor afvige meget fra parablen. Man siger at generaliseringsevnen er ringe. Desuden vil selv små forskydninger i inddataværdierne x_i bevirke store ændringer i det approksimerende polynomium. Det er derfor ikke antallet af data, men *strukturen i data*, der skal bestemme antallet af parametre.

Hvis man vælger for få parametre kan nettet ikke fungere som ønsket. Hvis man vælger for mange parametre vil nettet adaptere til støjen i data, og vil have en dårlig generaliseringsevne. Evnen til korrekt klassifikation af nye data vil i dette tilfælde være dårligere, end hvis færre parametre var valgt.

For neurale net er antallet af parametre ofte meget stort. Derfor er det især her vigtigt at vælge et passende antal neuroner. Desværre findes der ingen gode retningslinier for dette valg. Ofte er man nødt til at prøve sig frem. Vi skal senere diskutere nogle metoder til hvorledes et net kan reduceres.

6.3.3 Neurale net som feature ekstraktorer

Vi har tidligere set at de lagdelte netværk (uden tilbagekobling) blot implementerer et sæt af lineære beslutningsfunktioner. En anden måde at betragte en neurons funktionsmåde er som en korrelator. Ved lineær korrelation foretages jo også en sum af vægtede inddataværdier. Denne analogi holder netop når inddata betragtes som samplede værdier. Ved brug af neurale net på signal/billeddata er derfor naturligt at betragte neuronerne i input laget som feature ekstraktorer. I sådanne anvendelser er det som regel ikke rimeligt at netværket er fuldt forbundet. Alternativt bestemmes netforbindelserne svarende til støtten for ekstraktorene. Neuronerne i de følgende lag kan tænkes at foretage en klassifikation, eller en beregning af nogle ønskede mål.

En stor fordel ved neurale net er deres generalitet. Hvis man har et problem, som man ved kan løses, blot ikke hvordan, da kan brug af neurale net være fordelagtig. Ved efter oplæring af et net at undersøge de estimerede vægte, er det muligt at opnå god viden om hvilke feature ekstraktorer, filtre etc. som vil være hensigtsmæssige at anvende. Nettet kan således specificere hvorledes problemet kan løses. Ofte er det imidlertid vanskeligt at fortolke betydningen af de enkelte vægte.

6.4 Oplæring af neurale net

Som vi tidligere har set vil antallet af parametre i et neuralt net ofte være så stort, at det er nødvendigt at anvende en metode til estimation af parametrene. Der findes en række sådanne metoder. Vi skal her kun gennemgå en af de mest kendte, en simpel gradient descent (se Appendix B) kaldet *Back propagation algoritmen*. Fælles for alle metoder til oplæring af neurale net er, at et sæt af træningsdata er til rådighed. Ved gentagende anvendelse af nettet på et inddata, sammenligning af det producerede uddata med det forventede korrekte svar, og justering af vægtene i nettet sådan af fejlen mindskes, kan nettet oplæres. Lad $\{\vec{\mathbf{x}}, \vec{\mathbf{y}}\}$ være et sæt af M træningsdata, således at $\vec{\mathbf{y}}$ er det korrekte svar på inddata $\vec{\mathbf{x}}$. Lad nettets svar på $\vec{\mathbf{x}}$ være $\hat{\mathbf{y}}$. Da er fejlen givet ved den sædvanlige 2-norm givet ved:

$$E = \frac{1}{2}\sum_{k=1}^{M}(\vec{\mathbf{y}}_k - \hat{\mathbf{y}}_k)^t(\vec{\mathbf{y}}_k - \hat{\mathbf{y}}_k) \tag{6.4}$$

Centralt i det følgende, er bestemmelsen af en metode til justering af vægtene således at E minimeres.

Før man går i gang med udvikling af en optimeringsmetode til et problem, som det ovenstående, er det vigtigt at gøre sig klart dels problemets størrelse, dels karakteren af problemet. Som vi tidligere har set er antallet af vægte i et net ofte meget stort. Det rum, hvorover vi ønsker at minimere E er derfor tilsvarednde stort. Da hver neuron benytter en ulineær grænsefunktion vil E, betragtet som funktion af alle vægtene, i høj grad være ulineær. Funktionen E må derfor forventes at have mange lokale minima. Dette betyder at optimeringsproblemet er **meget** svært. Som diskuteret grundigere i Appendix B ved vi, at anvendelse af en simpel optimeringsmetode som fx. gradient descent med stor sandsynlighed vil resultere i bestemmelsen af et lokalt minima, som næppe er globalt. Problemstørrelsen indikerer at kraftigere (og langt mere beregningskrævende) metoder som fx. simuleret afkøling er nødvendige hvis sandsynligheden for at bestemme det globale minima skal øges. Dette er afgjort en mulighed. Imidlertid vil oplæringstiden ved anvendelse af sådanne metoder ofte være så astronomisk stor, at estimationen ikke vil kunne foretages i praksis. Derfor vil man ofte benytte en gradient descent alligevel.

6.4.1 Back-propagation algoritmen

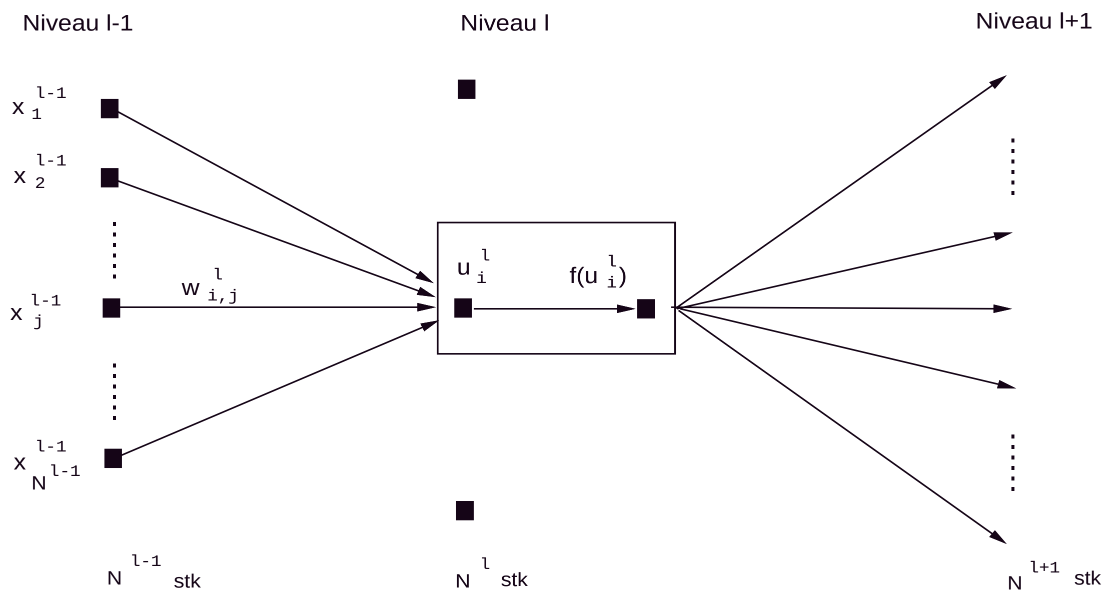
I udledningen af den gradient descent version, der kaldes back propagation algoritmen, skal vi studere hvorledes fejlen, som den måles ved uddataneuronen, kan propageres baglængs ned gennem nettet. Lad os først betragte neuronen $\mathbf{x}_i^l$ på det skjulte lag $l < L$ i nettet (hvor L er uddataniveauet). Forbindelselserne til og fra denne neuron er skitseret i figur 6.5.

Bemærk at notationen, jvf. figur 6.5, er at superscript angiver niveau i nettet, enkelt subscript angiver neuron nummer på det pågældende niveau, dobbeltsubscript på w_{ij}^l angiver vægten på forbindelsen fra knuden x_j^{l-1} til knuden x_i^l. Jævnfør tidligere definitioner har vi:

$$u_i^l = \sum_{k=1}^{N^{l-1}} w_{ik}^l \cdot x_k^{l-1} \qquad \text{og} \qquad x_i^l = f(u_i^l) \tag{6.5}$$

Ved differentiation fås let:

$$\frac{\partial u_i^l}{\partial x_j^{l-1}} = w_{ij}^l \qquad \text{og} \qquad \frac{\partial x_i^l}{\partial u_i^l} = f'(u_i^l) \tag{6.6}$$



Figur 6.5: Neuron nummer i på det skjulte lag nr $l < L$.

Vi skal benytte disse resultater og definitioner i det følgende. Bemærk, at den afledede f' af grænsefunktionen f optræder. Vi skal derfor antage at denne funktion er differentiabel.

Lad E være fejlen ved knuden x_i^l. Vi kan da skrive:

$$\frac{\partial E}{\partial w_{ij}^l} = \frac{\partial E}{\partial u_i^l} \cdot \frac{\partial u_i^l}{\partial w_{ij}^l} \tag{6.7}$$

$$= \frac{\partial E}{\partial u_i^l} \cdot \frac{\partial}{\partial w_{ij}^l} \left[\sum_{k=1}^{N^{l-1}} w_{ik}^l \cdot x_k^{l-1} \right] \tag{6.8}$$

$$= \frac{\partial E}{\partial u_i^l} \cdot x_j^{l-1} \tag{6.9}$$

$$= \delta_i^l \cdot x_j^{l-1} \tag{6.10}$$

Hvor $\delta_i^l = \frac{\partial E}{\partial u_i^l}$ er en definition. Vi har nu etableret en relation, der beskriver hvorledes fejlen E (differentielt) afhænger af vægten w_{ij}^l. Ønsker vi at mindske fejlen E, ved at justere vægten, da skal dette gøres med en størrelse, der er proportional med den negative gradient, dvs. $-\ \delta_i^l \cdot x_j^{l-1}$.

Det næste trin består i at etablere en relation mellem fejlen E og de knuder på niveau $l+1$ som x_j^l bidraget til. Bemærk først:

$$\delta_i^l = \frac{\partial E}{\partial u_i^l} = \frac{\partial E}{\partial x_i^l} \cdot \frac{\partial x_i^l}{\partial u_i^l} = \frac{\partial E}{\partial x_i^l} \cdot f'(u_i^l) \tag{6.11}$$

Ved brug af kædereglen på knuderne på niveau $l+1$ fås:

$$\frac{\partial E}{\partial x_i^l} = \sum_{k=1}^{N^{l+1}} \frac{\partial E}{\partial u_k^{l+1}} \cdot \frac{\partial u_k^{l+1}}{\partial x_i^l} = \sum_{k=1}^{N^{l+1}} (\delta_k^{l+1} \cdot w_{ik}^{l+1}) \tag{6.12}$$

Tilsammen får vi derfor:

$$\delta_i^l = f'(u_i^l) \cdot \sum_{k=1}^{N^{l+1}} \delta_k^{l+1} w_{ik}^{l+1} \tag{6.13}$$

Dette udtryk er centralt i back propagation algoritmen. Udtrykket beskriver hvorledes vi, givet kendskab til δ^{l+1} og w^{l+1} kan beregne δ^l, dvs, hvorledes størrelsen δ kan propageres baglængs ned gennem nettet. Opdateringsreglen for knude nummer i på niveau $l < L$ bliver således:

1. Beregn $\delta_i^l = f'(u_i^l) \cdot \sum_{k=1}^{N^{l+1}} \cdot \delta_k^{l+1} w_{ik}^{l+1}$
2. For alle vægte w_{ij}^l juster: $w_{ij}^l \leftarrow w_{ij}^l - \eta \cdot \delta_i^l \cdot x_j^{l-1}$

hvor η kaldes *indlæringsraten*. Andet punkt i skitsen svarer helt til en normal gradient descent, hvor hver parameter justeres med en størrelse, der er proportional med den negative gradient (der her er $-\frac{\partial E}{\partial w_{ij}^l}$).

For at kunne starte propagationen af fejlen E fra roden (output neuronerne) i nettet mangler vi at udlede et udtryk for δ_i^L på output niveauet L. Idet fejlen E (ved anvendelse på en enkelt inddata) på dette niveau er bestemt ved ligning (6.4) får vi af definitionen på δ at:

$$\delta_i^L = \frac{\partial E}{\partial u_i^L} \tag{6.14}$$

$$= \frac{\partial E}{\partial \hat{y}_i^L} \cdot \frac{\partial \hat{y}_i^L}{\partial u_i^L} \tag{6.15}$$

$$= \frac{\partial}{\partial \hat{y}_i^L} \left[\frac{1}{2} \sum_{k=1}^{N^L} (y_k - \hat{y}_k)^2 \right] \cdot f'(u_i^L) \tag{6.16}$$

$$= (\hat{y}_i - y_i) \cdot f'(u_i^L) \tag{6.17}$$

Vi har nu gennemgået de enkelte trin i back propagation algoritmen, men mangler stadig at konstruere en egentlig træningsalgoritme. Antag at der er givet M stk. samhørende inddataværdier x_t og korrekte uddataværdier y_t. En simpel skitse er da:

```
for t = 1 to M do begin
   y-hat_t = NET(x_t);
   for l = L to 1 do begin
       for i = 1 to N^l do begin
```

```
                    Beregn delta_i^l
                    for j = 1 to N^{l-1} do begin
                        Juster vaegt w_{ij}^l
                    end
                end
            end
        end
```

I ovenstående skitse foretages optimeringen efter hver inddatavektor. Dette svarer ikke helt til en sædvanlig gradient descent, hvor korrektionen af vægtene først foretages efter udregning af summen af samtlige fejl. Ovenstående metode kaldes derfor ofte for *Noisy gradient descent*. Ofte er det nødvendigt at iterere ovenstående skitse mange gange før den samlede fejl bliver acceptabelt lille. Det er altså ikke tilstrækkeligt kun at præsentere hvert datasæt $(\mathbf{x}, \mathbf{y})$ én gang for nettet. Metoden er, som tidligere omtalt, meget simpel, og der er derfor også udviklet en lang række forbedringer af denne. Det vil føre for vidt her at diskutere disse varianter.

Et praktisk problem ved anvendelse af neurale net er, at antallet træningsdata bør være stor (gerne således at antal inddataværdier overstiger antal vægte). Ofte er der imidlertid ikke så mange træningsdata til rådighed. En brugbar (og meget effektiv) metode er at tilføje træningsdata fremkommet af de oprindelige ved kunstigt at tilføre disse støj. Disse nye data skal have samme korrekte uddata som de støjfri. Ud over at øge træningssættet er fordelen ved metode, at nettet oplæres til at abstrahere fra støj i data. Dette er næsten altid ønskeligt.

6.4.2 Udtynding af neurale net

Efter oplæring af et net kan det være hensigtsmæssigt at undersøge om visse af forbindelserne kan undværes. Hvis et stort antal forbindelser kan slettes, vil videre oplæring af nettet blive nemmere. Ved implementation af nettet på en sædvanlig sekvential maskine vil en reduktion af netstørrelsen betyde et tilsvarende speed-up. Udtynding af et net kan mest enkelt foretages ved at fjerne alle forbindelser hvor den tilknyttede vægt er under en vis lille tærskelværdi. Denne metode kræver at tærskelvædien adapteres til størrelsesordenen af de andre vægte (der fører til knuden). Denne metode har imidlertid vist sig at fungere dårligt i praksis. Alternativt er det muligt, under oplæringen af nettet, at normere vægtene, således at en fast konstant tærskelværdi kan bruges. Der er i de senere år forsket meget i bedre metoder til udtynding af neurale net. Det vil føre for vidt her at diskutere disse metoder.

6.4.3 Praktisk brug af neurale net

I de senere år er der fremkommet arkitekturer, der direkte i silicium, implementerer små neurale net. Disse er, pga. nettets simple beregningsmåde, meget hurtige. Problemet ved neurale net er derfor primært, at oplære nettet, herunder at bestemme antal neuroner, samt de nødvendige forbindelser. Erfaringen tyder på, at hvis problemstørrelsen (og dermed netstørrelsen) holdes lille, da er det muligt at konstruere effektive neurale net. Oplæringstiden er imidlertid stadig stor. Når udviklingstiden tages i betragtning er der derfor et stort spørgsmål om klassifikation ved neurale net er fordelagtige frem for brugen af statistiske metoder.

6.5 Neurale net contra diskriminantanalyse

Fordelene ved neurale net er, at de er generelle, lette at forstå, og at de kan modellere enhver funktion, når blot antallet af neuroner er stort nok. Hvis man ingen viden har om hvilke features, der er gode, hvordan klasserne stukturerer sig i featurerummet, eller hvilken form klasserne har, da vil neurale net ofte være anvendelige. I bedste fald vil de kunne angive de optimale features. Ved implementation i silicium vil neurale net være meget effektive.

Ulempen ved neurale net er, at det kan være meget svært at vælge antallet af neuroner i de skjulte lag optimalt. Oplæringstiden for selv små neurale net er ofte meget stor. Ved anvendelse af simple optimeringsmetoder (som back propagation algoritmen) er sandsynligheden for at ende i et lokalt minimum stor (dette afhænger naturligvis af problemets karakter).

Samenlignet med statistiske klassifikationsmetoder, som diskriminantanalysen, er en væsenlig forskel, at de neurale net ikke forudsætter at features er bestemt på forhånd. Ligeledes er det intet krav at klyngerne i data kendes og deres middelværdi og kovariansmatrice (e.lign) er estimeret. Således kan man sige at disse processer, i neurale net, er kogt sammen til ét meget stort optimeringsproblem. Hvis klasserne i featurerummet ligger tæt og disse har en krum form (som fx en normalfordeling) vil antallet af nødvendige neuroner være stort. Her vil diskriminantfunktionerne, som jo er udledt fra en parametrisk estimation af tæthedsfordelingen for en klasse, kunne modellere observationerne med langt færre parametre.

6.6 Opgaver

1. Lad (x, y, ω_i), hvor x og y er to features og ω_i er den kendte klasse, angive et sæt af træningsdata hvorpå et neuralt feed-forward lagdelt netværk tænkes optrænet. Lad data være givet som i skemaet:

x	y	ω	x	y	ω
4	9	1	2	5	4
6	9	2	3	5	1
7	9	2	6	5	3
3	8	1	7	5	3
4	8	1	8	5	3
6	8	2	3	4	4
7	8	2	4	4	4
3	7	1	5	4	4
4	7	1	6	4	5
5	7	1	7	4	5
6	7	3	3	3	4
7	7	2	4	3	4
3	6	1	5	3	4
4	6	1	6	3	5
6	6	3	7	3	5
7	6	2	4	2	4

Spørgsmål 1: Er klasserne lineært separable?

1) Ja
2) Nej, klasse 1 og 2 er ikke lineært separable
3) Nej, klasse 1 og 3 er ikke lineært separable
4) Nej, klasse 2 og 3 er ikke lineært separable
5) Nej, klasse 1 og 4 er ikke lineært separable
6) Nej, klasse 3 og 5 er ikke lineært separable
7) Nej, klasse 4 og 5 er ikke lineært separable
8) Nej, ingen af klaserne er ikke lineært separable
9) Ved ikke

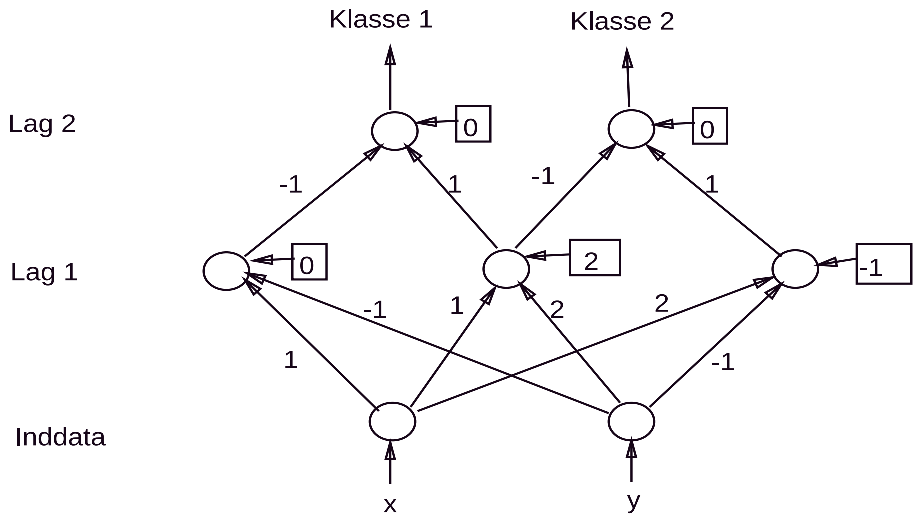
2. Implementer back propagation algoritmen på et simpelt 2 lags feed-forward net med to inputneuroner. Lad der være 5 output neuroner. Netværket skal optrænes på de data der er anført i opgave 1. Hvor mange neuroner skal mellemlaget indeholde? Gennem hvor mange iterationer skal vægtene justeres før fixpunktet er fundet?

3. Antag at der er givet tre prototyper A:(2,3); B:(0,-1); og C:(4,1), og antag at alle observationer, der med sikkerhed kan henføres til en af de tre prototyper, har en maksimal afstand fra prototypen på 1.0, idet maksimumsnormen benyttes. Er klasserne lineært seperable? Alle observationer, der ligger uden for klyngerne henføres til en særlig prototype F (for Fejl). Design et neuralt net, der kan klassificere enhver observation til en af de fire prototyper.

4. Betragt nedenstående 2-lags net med to output neuroner $Klasse_1$ og $Klasse_2$, og tre neuroner i det første lag. Lad inddata til nettet være x og y, der begge antages heltallige. Lad endvidere vægtene i nette være givet som vist på nedenstående figur,

hvor biasværdierne er angivet i firkanter. Antag at der benyttes en grænsefunktion givet ved:

$$f(t) \;=\; \begin{cases} 1 & \text{hvis} \quad t \geq 0 \\ 0 & \text{ellers} \end{cases}$$



Figur 6.6:

Beskriv, gerne vha. en tegning, funktionen af det neurale net.

Bibliografi

[1] S. Banks: *Signal processing, Image Processing, and Pattern Recognition*; Prentice Hall 1990.

[2] A. Blake, A. Zisserman: *Visual Reconstruction*; The MIT Press 1987.

[3] R. O. Duda, P. E. Hart: *Pattern Classification and Scene Analysis*; John Wiley 1973.

[4] L. Fausett: *Fundamentals of Neural Networks, architectures, algorithms, and applications*; Prentice Hall 1994.

[5] K. Fukunaga: *Introduction to Statistical Pattern Recognition*, 2. ed.; Academic Press 1990.

[6] A. Hald: *Statistiske Metoder*; Akademisk Forlag, 1977.

[7] P.J.M. van Laarhoven, E.H.L. Aarts: *Simulated Annealing: Theory and Applications*; D. Reidel Publishing 1987.

[8] E. Másson, Y.-J. Wang: *Introduction to Computation and Learning in Neural networks*; DIKU-report nr. 89/3

[9] W. H. Press et.al.: *Numerical Recipes in C*, 2. ed; Cambridge Press 1992.

[10] J. Rissanen: *Stochastic Complexity in Statistical Inquiry*; World Scientific 1989.

[11] P. J. Rosseeuw, A. M. Leroy: *Robust Regression and Outlier Detection*; John Wiley 1987.

[12] Schaum's outline series: *Theory and problems of Matrices*; McGraw-Hill 1974

[13] A. H. Andersen: *Forelæsningsnoter*; Afdeling for Teoretisk Statistik, Matematisk Institut, Aarhus Universitet, 1976.

[14] J. T. Tou, R. C. Gonzalez: *Pattern Recognition Principles*; Addison-Wesley, 1994.

Appendix A

Grundlæggende Lineær Algebra

Dette appendix omhandler nogle få væsentlige elementer af lineær algebra som kan være til hjælp i læsningen af den øvrige del af noterne, men kan ikke erstatte egenligt undervisningsmateriale. Hovedparten af kapitlet burde være velkendt fra den grundlæggende matematik. Afsnittet om løsning af lineære ligningssystemer er medtaget for at give en idé om nogle af de i praksis meget benyttede metoder. De beregningstekniske overvejelser diskuteres ikke. Tilsvarende udelades egentlige algoritmeskitser, idet dette bedre hører til i et kursus om numerisk analyse. Det nok væsentligste, for en fuld forståelse af resten af noterne, er at læseren bliver fortrolig med egenværdi-dekomposition af matricer, samt af egenskaberne ved positivt definitte matricer.

A.1 Vektorrum, indre produkt, norm, basis

Nedenstående defineres vektorrum $(\mathcal{V}, +, *)$ over de reelle tals legeme $\mathcal{R}$. Andre legemer, f.eks. de komplekse tals legeme $\mathcal{C}$ kunne være benyttet i stedet. Som $\mathcal{V}$ vil vi tænke på $\mathcal{R}^n$ eller $\mathcal{C}^n$. En ikke tom mængde $\mathcal{V}$ kaldes et *vektorrum* over $\mathcal{R}$ hvis der findes to afbildninger:

$$\begin{array}{lll} \text{addition} & \mathcal{V} \times \mathcal{V} \to \mathcal{V} & : (\mathbf{x}, \mathbf{y}) \to \mathbf{x} + \mathbf{y} \\ \text{skalar multiplikation} & \mathcal{R} \times \mathcal{V} \to \mathcal{V} & : (\lambda, \mathbf{x}) \to \lambda \mathbf{x} \end{array}$$

således at nedenstående betingelser er opfyldt for alle $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{V}$, og for alle $\mu, \lambda \in \mathcal{R}$:

1. $\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$
2. $(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z})$
3. $\exists\, \mathbf{0} : \mathbf{x} + \mathbf{0} = \mathbf{x}$
4. $\exists - \mathbf{x} : \mathbf{x} + (-\mathbf{x}) = \mathbf{0}$
5. $\lambda(\mathbf{x} + \mathbf{y}) = \lambda\mathbf{x} + \lambda\mathbf{y}$
6. $(\lambda + \mu)\mathbf{x} = \lambda\mathbf{x} + \mu\mathbf{x}$
7. $(\lambda\mu)\mathbf{x} = \lambda(\mu\mathbf{x})$
8. $1\mathbf{x} = \mathbf{x}$

Et element $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in \mathcal{V}$ kaldes en vektor, og n kaldes dimensionen af vektorrummet. En *basis* for et vektorrum $\mathcal{V}$ (over $\mathcal{R}$) er en mængde af vektorer $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\} \in \mathcal{V}$ således at enhver vektor $\mathbf{x}$ har en entydig repræsentation:

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

hvor $\alpha_i \in \mathcal{R}$.

En *norm* på et vektorrum $\mathcal{V}$ er en afbildning $||\cdot|| : \mathcal{V} \to \mathcal{R}^+$, der opfylder at:

$$\begin{array}{rcll} ||\mathbf{x}|| & = & 0 & \text{hvis og kun hvis } \mathbf{x} = \mathbf{0} \\ ||\alpha\mathbf{x}|| & = & |\alpha|||\mathbf{x}|| & \text{for alle } \alpha \in \mathcal{R},\ \mathbf{x} \in \mathcal{V} \\ ||\mathbf{x}+\mathbf{y}|| & \leq & ||\mathbf{x}|| + ||\mathbf{y}|| & \text{for alle } \mathbf{x}, \mathbf{y} \in \mathcal{V} \end{array}$$

Sidstnævnte krav kaldes *trekantsuligheden*. Hvis der findes $\mathbf{x} \neq 0$ for hvilke $||\mathbf{x}|| = 0$ kaldes afbildningen for en *semi-norm*.

Ved et *indre produkt* i et vektorrum $\mathcal{V}$ over $\mathcal{C}$ forstås en afbildning $\mathcal{V} \times \mathcal{V} \to \mathcal{R}$ sådan at:

$$\begin{array}{lrcll} 1. & <\mathbf{x},\mathbf{x}> & \geq & 0 & \text{for alle } \mathbf{x} \in \mathcal{V} \\ 2. & <\mathbf{x},\mathbf{x}> & = & 0 & \text{hvis og kun hvis } \mathbf{x} = 0 \\ 3. & <\mathbf{x},\mathbf{y}> & = & \overline{<\mathbf{y},\mathbf{x}>} & \text{for alle } \mathbf{x}, \mathbf{y} \in \mathcal{V} \\ 4. & <\lambda\mathbf{x} + \mu\mathbf{y}, \mathbf{z}> & = & \lambda <\mathbf{x},\mathbf{z}> + \mu <\mathbf{y},\mathbf{z}> & \text{for alle } \lambda, \mu \in \mathcal{C} \text{ og for alle } \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{V} \end{array}$$

hvor $\overline{x}$ angiver den komplekst konjugerede af x. Hvis vektorrummet er defineret over $\mathcal{R}$ bortfalder den komplekse konjugering. Hvis det andet krav ikke er opfyldt kaldes afbildningen for et *semi-indre produkt*. I vektorrummet $\mathcal{R}^n$ over $\mathcal{R}$ er *det sædvanlige indre produkt* defineret ved:

$$<\mathbf{x},\mathbf{y}> = (x_1, x_2, \ldots, x_n) \cdot (y_1, y_2, \ldots, y_n)^t = x_1y_1 + x_2y_2 + \ldots + x_ny_n = \sum_{i=1}^{n} x_iy_i$$

Hvis et vektorrum $\mathcal{V}$ er udstyret med et indre produkt da vil afbildningen $\mathbf{x} \to \sqrt{<\mathbf{x},\mathbf{x}>}$ definere en norm. Tre ofte benyttede normer på $\mathcal{R}^n$ er:

$$\begin{array}{rcl} \mathbf{x} = (x_1, x_2, \ldots, x_n) \to ||\mathbf{x}|| & = & max\,\{|x_1|, |x_2|, \ldots, |x_n|\} \\ \mathbf{x} = (x_1, x_2, \ldots, x_n) \to ||\mathbf{x}|| & = & |x_1| + |x_2| + \cdots + |x_n| \\ \mathbf{x} = (x_1, x_2, \ldots, x_n) \to ||\mathbf{x}|| & = & \sqrt{x_1^2 + x_2^2 + \cdots x_n^2} \end{array}$$

Den første norm kaldes *maksimumsnormen* (eller l_∞-normen). Den anden norm kaldes l_1-normen. Den tredie norm kaldes den *euclidiske norm* (eller l_2-normen).

Hvis der om en basis $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ for et vektorrum $\mathcal{V}$ gælder at $||\mathbf{e}_i|| = 1$ for alle i, siges basis vektorene at være *normeret*. Hvis der yderligere gælder at det indre produkt $<\mathbf{e}_i, \mathbf{e}_j> = 0$ for alle $i \neq j$ kaldes basen for *orthonormal*. Benyttes Diracs delta-funktion:

$$\delta(i,j) = \begin{cases} 1 & \text{hvis } i = j \\ 0 & \text{hvis } i \neq j \end{cases}$$

kan betingelsen for orthonormalitet n skrives som: $<\mathbf{e}_i, \mathbf{e}_j> = \delta(i,j)$.

A.2 Koordinater, matricer

Lad $\mathcal{V}$ være et vektorrum med indre produkt, og lad $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ være en ortonormal basis. Givet en vektor $\mathbf{x}$ definerer $x_i =< \mathbf{x}, \mathbf{e}_i >$ længden af *projektionen* af $\mathbf{x}$ på $\mathbf{e}_i$. Projektionen selv kan skrives som $x_i\mathbf{e}_i$. Givet en orthonormal basis kan $\mathbf{x}$ således opskrives på entydig måde:

$$\mathbf{x} = \sum_{i=1}^{n} x_i \mathbf{e}_i$$

Koordinaterne for vektoren $\mathbf{x}$ (underforstået mht. basen $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$) skrives $(x_1, x_2, \ldots, x_n)^t$. Udtrykt i den orthonormale basis har basisvektorene koordinaterne:

$$\begin{array}{rcl}
\mathbf{e}_1 & = & (1, 0, 0, \ldots, 0)^t \\
\mathbf{e}_2 & = & (0, 1, 0, \ldots, 0)^t \\
\mathbf{e}_3 & = & (0, 0, 1, \ldots, 0)^t \\
\vdots & \vdots & \qquad\qquad \vdots \\
\mathbf{e}_n & = & (0, 0, 0, \ldots, 1)^t
\end{array}$$

En *matrice* er en opstilling af m vektorer søjle for søjle. Hvis hver vektor har n elementer vil matricen have dimensionen $n \times m$. Matricen bestående af ovenstående enhedsvektorer kaldes en *enhedsmatrice*.

$$\mathbf{I} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & & \vdots \\ & & \ddots & 0 \\ 0 & \ldots & 0 & 1 \end{pmatrix}$$

Enhedsmatricen er en *diagonalmatrice* hvor alle elementer i diagonalen har værdi lig 1. En diagonalmatrice med elementer $d_1, d_2, \ldots, d_n$ skrives ofte $diag(d_1, d_2, \ldots, d_n)$.

En nedre hhv. øvre trekantsmatrice har lutter nuller over hhv. under diagonalen.

$$\mathbf{L} = \begin{bmatrix} l_{11} & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 \\ l_{31} & l_{32} & l_{33} & 0 \\ l_{41} & l_{42} & l_{43} & l_{44} \end{bmatrix} \quad \text{og} \quad \mathbf{U} = \begin{bmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ 0 & u_{22} & u_{23} & u_{24} \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{bmatrix}$$

En diagonalmatrice er således såvel en øvre trekantsmatrice som en nedre trekantsmatrice.

Lad $(\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n)$ være et sæt af n vektorer, og lad $\mathbf{X}$ være matricen med søjler givet ved de n vektorer, dvs. $\mathbf{X} = (x_{ij})$ hvor $x_{ij} = \mathbf{x}_i^j$. Vektorene $(\mathbf{x}^j)$ siges at være *lineært uafhængige* hvis der gælder ingen af disse kan skrives som en linearkombination af de øvrige. Dette betyder at den eneste vektor $\mathbf{y}$ hvorom $\mathbf{X}\mathbf{y} = \mathbf{0}$ er vektoren $\mathbf{y} = \mathbf{0}$. Hvis vektorene $\mathbf{x}^j$ er lineært uafhængige udgør de en ny basis. En given vektor $\mathbf{y}$, har naturligvis forskellige koordinater alt efter hvilken basis der benyttes. Transformationer mellem forskellige koordinatsystemer (baser) er derfor en central operation. Lad koordinaterne for vektoren $\mathbf{y}$ være $(y_1, y_2, \ldots, y_n)^t$. I koordinatsystemet givet ved basisvektorene $(\mathbf{x}^j)$ vil vektoren have koordinaterne $\tilde{y_i}$ hvor:

$$\tilde{y_i} = \sum_{j=1}^{n} x_{ji} y_j$$

Mere kompakt kan dette udtrykkes på matrixform ved: $\tilde{\mathbf{y}} = \mathbf{X^t y}$. Matricen $\mathbf{X^t}$ beskriver således en vektorafbildning. Billedet $\tilde{\mathbf{e}}_i$ af $\mathbf{e}_i$ er givet ved $\tilde{\mathbf{e}_i} = \mathbf{X^t e}_i$. Matricen $\mathbf{X}$ kaldes *singulær* hvis der eksisterer to vektorer $\mathbf{y}_1 \neq \mathbf{y}_2$ således at billederne af disse vektorer er ens, dvs. hvis $\mathbf{X}\mathbf{y}_1 = \mathbf{X}\mathbf{y}_2$. Søjlerne i en singulær matrice er altså lineært afhængige.

Lad der være givet en $n \times m$ matrice $\mathbf{X}$. Antallet af lineært uafhængige rækker hhv. søjler af $\mathbf{X}$ kaldes for *rækkerangen* hhv. *søjlerangen* af $\mathbf{X}$. Det kan vises at *rækkerangen* og *søjlerangen* er ens, hvorved vi kan nøjes med at tale om *rangen* af $\mathbf{X}$. Hvis dimensionen n af det vektorrum der afbildes fra er lig dimensionen m af det vektorrum der afbildes på er ens (dvs. $m = n$) kaldes matricen for *kvadratisk*. Hvis rangen af en kvadratisk matrice er lig dimensionen af matricen kaldes denne for *regulær*.

Summen $\mathbf{X} + \mathbf{Y}$ af to matricer $\mathbf{X}$ og $\mathbf{Y}$ af samme dimension har elementer givet ved summen $x_{ij} + y_{ij}$ af elementerne i de to matricer. Produktet af en matrice (x_{ij}) med en skalar λ er en matrice med elementer λx_{ij}. Det verificeres let at mængden af matricer med reelle/complekse elementer ved disse kompositioner er organiseret som et vektorrum over $\mathcal{R}$ eller $\mathcal{C}$. Produktet af en $n \times m$ matrice $\mathbf{X} = (x_{ij})$ med en $m \times p$ matrice $\mathbf{Y} = (y_{jk})$ er defineres ved $n \times p$ matricen $\mathbf{XY}$:

$$\mathbf{XY} \ = \ (x_{ik}) \ = \ (\sum_{j=1}^{m} x_{ij} y_{jk})$$

Matricemultiplikation er ikke generelt kommutativ, dvs: $\mathbf{XY} \neq \mathbf{YX}$. Derimod er matricemultiplikation såvel associativ som distributiv.

Hvis der om en kvadratisk matrice $\mathbf{X}$ gælder at $\mathbf{XX} = \mathbf{X}^2 = \mathbf{X}$ siges $\mathbf{X}$ at være *idempotent*. Hvis $\mathbf{X}^2 = \mathbf{I}$ kaldes $\mathbf{X}$ for *involutorisk*. Hvis der eksisterer et p således at $\mathbf{X}^p = \mathbf{O}$ siges $\mathbf{X}$ at være *nilpotent*. Hvis der eksisterer en matrice $\mathbf{Y}$ således at $\mathbf{XY} = \mathbf{I}$ kaldes $\mathbf{Y}$ den *inverse* til $\mathbf{X}$, dvs. $\mathbf{Y} = \mathbf{X}^{-1}$. Alle regulære matricer har en invers. Den inverse til diagonalmatricen $diag(d_1, d_2, \ldots, d_n)$ er diagonalmatricen $diag(1/d_1, 1/d_2, \ldots, 1/d_n)$.

Den *transponerede* af en matrice $\mathbf{X} = (x_{ij})$, er en matrice $\mathbf{X}^t$, hvor elementerne i den j'te række og i'te søjle er lig x_{ij}, dvs. at $(x_{ij})^t = (x_{ji})$. Der gælder at:

$$\begin{array}{ll} (\mathbf{X}^t)^t = \mathbf{X} & (\lambda\mathbf{X})^t = \lambda\mathbf{X}^t \\ (\mathbf{X}+\mathbf{Y})^t = \mathbf{X}^t + \mathbf{Y}^t & (\mathbf{XY})^t = \mathbf{Y}^t\mathbf{X}^t \end{array}$$

En matrice kaldes *symmetrisk* hvis $\mathbf{X}^t = \mathbf{X}$. Hvis $\mathbf{X}^t = -\mathbf{X}$ kaldes $\mathbf{X}$ for *skævsymmetrisk*. Enhver kvadratisk matrice $\mathbf{A}$ kan dekomponere som summen af en symmetrisk matrice $\mathbf{A}^+$ og en skævsymmetrisk matrice $\mathbf{A}^-$.

$$\mathbf{A} \ = \ \mathbf{A}^+ \ + \ \mathbf{A}^- \ = \ \frac{1}{2}(\mathbf{A} + \mathbf{A}^t) \ + \ \frac{1}{2}(\mathbf{A} - \mathbf{A}^t)$$

En reel symmetrisk (herunder kvadratisk) matrice $\mathbf{X}$ kaldes *positiv definit* hvis der eksisterer en ikke-singulær matrice $\mathbf{C}$ således at $\mathbf{X} = \mathbf{C}^t\mathbf{C}$. En kvadratisk matrice $\mathbf{X}$ kaldes *ortogonal* hvis der gælder at $\mathbf{X}^t\mathbf{X} = \mathbf{X}\mathbf{X}^t = \mathbf{I}$, dvs. hvis $\mathbf{X}^t = \mathbf{X}^{-1}$. Jvf. det tidligere er søjlevektorene for en ortogonal matrice ortonormale.

For en kompleks matrice $\mathbf{X}$ er den *komplekst konjugerede* matrice $\overline{\mathbf{X}}$ bestemt ved konjugering af elementerne i $\mathbf{X}$. Der gælder at:

$$\begin{array}{ll} \overline{(\overline{\mathbf{X}})} = \mathbf{X} & \overline{(\lambda\mathbf{X})} = \overline{\lambda}\,\overline{\mathbf{X}} \\ \overline{(\mathbf{X}+\mathbf{Y})} = \overline{\mathbf{X}} + \overline{\mathbf{Y}} & \overline{(\mathbf{X}\mathbf{Y})} = \overline{\mathbf{Y}}\,\overline{\mathbf{X}} \end{array}$$

Den *adjungerede* $\mathbf{X}^*$ af en matrice $\mathbf{X}$ er defineret ved $\mathbf{X}^* = \overline{\mathbf{X}}^t = \overline{\mathbf{X}^t}$. Hvis matricen har reelle elementer gælder $\mathbf{X}^* = \mathbf{X}^t$. Der gælder:

$$\begin{array}{ll} (\mathbf{X}^*)^* = \mathbf{X} & (\lambda\mathbf{X})^* = \overline{\lambda}\mathbf{X}^* \\ (\mathbf{X}+\mathbf{Y})^* = \mathbf{X}^* + \mathbf{Y}^* & (\mathbf{X}\mathbf{Y})^* = \mathbf{Y}^*\mathbf{X}^* \end{array}$$

Hvis $\mathbf{X}$ er regulær gælder der at $(\mathbf{X}^*)^{-1} = (\mathbf{X}^{-1})^*$. Hvis $\mathbf{X}^* = \mathbf{X}$ kaldes matricen for *Hermitisk*. Hvis $\mathbf{X}^* = -\mathbf{X}$ kaldes matricen for *anti-Hermitisk*. Hvis $\mathbf{X}^* = \mathbf{X}^{-1}$ kaldes matricen for *unitær*.

Ved *Sporet* $Tr(\mathbf{X})$ af en $n \times n$ matrice $\mathbf{X}$ forstås summen $\sum_i \mathbf{x}_{ii}$ af diagonalelementerne. Bemærk, at det indre produkt $\mathbf{x}^t\mathbf{y}$ af af to vektorer $\mathbf{x}$ og $\mathbf{y}$ er lig sporet af det ydre produkt $\mathbf{x}\mathbf{y}^t$

Ved *determinanten* af en $n \times n$ matrice $\mathbf{X}$ forstås tallet:

$$det\,\mathbf{X} = |\mathbf{X}| = \sum_{p \in S_n} sign(p) x_{p(1)1} \cdots x_{p(n)n}$$

hvor der summeres over samtlige $n!$ produkter af n elementer fra matricen bestående af et element fra hver søjle, således at hver række er repræsenteret en gang. S_n angiver samtlige permutationer af de n indices, og $sign(p)$ angiver om permutationen er lige (+1) eller ulige (-1). $p(i)$ angiver det i'te element af den p'te permutation. Determinanten er $\neq 0$ hvis og kun hvis matricen er regulær. Hvis $det\,\mathbf{X} \neq 0$ har matricen en invers. For to $n \times n$ matricer $\mathbf{X}$ og $\mathbf{Y}$ gælder at $det\,(\mathbf{X}\mathbf{Y}) = det\,\mathbf{X}\; det\,\mathbf{Y}$, og at $det\,\mathbf{X}\; det\,\mathbf{X}^{-1} = 1$. Hvis $\mathbf{X}$ er ortogonal eller unitær gælder at $|det\,\mathbf{X}| = 1$.

A.3 Egenværdier, egenvektorer, konditionstal

Lad $\mathbf{A}$ være en $n \times n$ matrice og lad $\mathbf{x} \neq \mathbf{0}$ være en vektor af dimension n. Hvis der gælder at:

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

hvor λ er en skalar, da kaldes $\mathbf{x}$ for en *egenvektor* til $\mathbf{A}$, og λ kaldes den til egenvektoren hørende *egenværdi*. Der findes netop n (ikke nødvendigvis forskellige) egenværdier for $\mathbf{A}$.

Fra ovenstående ligning har vi at:

$$\lambda\mathbf{x} - \mathbf{A}\mathbf{x} = (\lambda\mathbf{I} - \mathbf{A})\mathbf{x} = \begin{bmatrix} \lambda - a_{11} & -a_{12} & \cdots & -a_{1n} \\ -a_{21} & \lambda - a_{22} & \cdots & -a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ -a_{n1} & -a_{n2} & \cdots & \lambda - a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = 0$$

Denne ligning har en ikke triviel løsning hvis og kun hvis $|\lambda\mathbf{I} - \mathbf{A}| = 0$. Ved udskrivning af determinanten fås et n'te grads polynomium i λ. Dette polynomie kaldes *det karakteristiske polynomium.* Egenværdierne er bestemt ved rødderne i det karakteristiske polynomium. Hvis der blandt de n egenværdier findes en værdi der optræder p gange siges denne at have *multiplicitet p*. Det kan vises at rangen af $\mathbf{A}$ er lig antallet af egenværdier (talt med multiplicitet), der er forskellige fra 0.

Det kan vises, at alle egenværdier for en reel symmetrisk matrice er positive hvis og kun hvis matricen er positiv definit. Omvendt gælder at egenværdierne for en skævsymmetrisk matrice er lig 0 eller er rent imaginære. Hvis $\mathbf{A}$ er reel og symmetrisk, da findes en ortogonal matrice $\mathbf{P}$ således at:

$$\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{D} = diag(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

hvor $\lambda_1, \lambda_2, \ldots, \lambda_n$ er egenværdierne for $\mathbf{A}$. Det kan også vises at søjlerne i $\mathbf{P}$ er bestemt ved egenvektorerne til $\mathbf{A}$. Hvis $\mathbf{A}$ er reel, men ikke er symmetrisk, vil sættet af egenvektorer ikke i almindelighed være ortogonale.

Da $\mathbf{P}$ er ortogonal ($\mathbf{P}^{-1} = \mathbf{P}^t$) ses af ovenstående udstryk at $\mathbf{A}^{-1} = \mathbf{P}^t\mathbf{D}^{-1}\mathbf{P}$. Invertering af en matrice er derfor en enkel operation når blot egenvektorer og egenværdier er kendt. Forudsætningen for inverteringen er at alle egenværdier er forskellige fra 0. Hvis matricen $\mathbf{A}$ er tæt på at være singulær, dvs. hvis en eller flere af egenværdierne er tæt på 0 kan det forventes at bestemmelsen af $\mathbf{A}^{-1}$ vil være unøjagtig. Dette skyldes de numeriske problemer ved regning på en datamat med endelig præcision. Bemærk at en multiplikation af alle elementer i matricen $\mathbf{A}$ med en skalar vil bevirke at alle egenværdier bliver multipliceret med samme faktor. Derfor er den numeriske mindste egenvædi et dårligt mål for hvor tæt matricen $\mathbf{A}$ er på at være singulær. I stedet bruges forholdet mellem den største og mindste egenværdi:

$$\kappa = \frac{\max\{\lambda_1, \lambda_2, \ldots, \lambda_n\}}{\min\{\lambda_1, \lambda_2, \ldots, \lambda_n\}}$$

κ kaldes *konditionstallet* for matricen. En lille værdi af κ angiver at matricen er godt konditioneret, og at en invertering vil være numerisk stabil. Er κ omvendt stor, er matricen dårligt konditioneret, og der grund til bekymring.

Antallet af egenværdier, der er forskellige fra 0, angiver dimensionen af det underrum af $\mathcal{R}^n$ hvori en vektor vil blive afbildet. Tilsvarende vil den relative størrelse af en egenværdi indikere betydningen af underrummet svarende til egenvektoren. *Energien* E af en lineær transformation kan defineres ved summen af de numeriske egenværdier.

$$E = |\lambda_1| + |\lambda_2| + \cdots + |\lambda_n|$$

Det procentuelle informationsindholdet I_i af den i'te egenvektor kan da defineres ved

$$I_i = \frac{|\lambda_1|}{100 \cdot E} \%$$

Hvis egenvædierne antages sorteres i (numerisk) voksende rækkefølge $|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$ og $\mu_j = \sum_{i=1}^{j} I_i$ siges at μ_j % af energien ved transformationen er koncentreret i et j-dimensionalt underrum.

To matricer **A** og **B** kaldes *ækvivalente* hvis der findes en ikke-singulær matrice **R** således at:

$$\mathbf{B} = \mathbf{R}^{-1}\mathbf{A}\mathbf{R}$$

To ækvivalente matricer har samme egenværdier. Hvis **b** er egenvektoren for **B** svarende til egenværdien λ, da vil **Rb** være egenvektoren svarende til egenværdien λ for **A**. Hvis en $n \times n$ matrice **A** har rang n, så vil **A** være ækvivalent med en diagonalmatrice. Hvis **A** har rang $< n$ vil **A** ikke være ækvivalent med en diagonalmatrice. Alle kvadratiske matricer kan imidlertid vises at være ækvivalente med en trekantsmatrice hvis diagonalelementer er lig egenværdierne for **A**. Hvis egenværdierne er reelle eksisterer en ortogonal matrice **P** således at $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{P}^t\mathbf{A}\mathbf{P}$ er en trekantsmatrice med egenværdierne for **A** som diagonalelementer.

A.4 Lineære ligningssystemer

Lad **X** være en reel $k \times n$ matrice, og **y** en vektor med k elementer. Da **X** beskriver en transformation fra et n-dimensionalt vektorrum $\mathcal{R}^n$ på et k-dimensionalt vektorrum $\mathcal{R}^k$ er der nærliggende at spørge hviken vektor i $\mathcal{R}^n$ der afbildes i **y**. Dette svarer til ligningen:

$$\mathbf{y} = \mathbf{X}\mathbf{a}$$

Ved udskrivning ses (matrix-) ligningen at indeholde k sædvanlige lineære ligninger i n variable $(a_1, a_2, \ldots, a_n)$. Dette system af ligninger kaldes *overbestemt* hvis der er flere ligninger end ubekendte, dvs. hvis $k > n$. Hvis $k < n$ kaldes systemet *underbestemt*. Hvis $k = n$ kaldes systemet *kvadratisk*. Kun i dette tilfælde er der mulighed for at systemet har en entydig løsning, dvs. et sæt af vædier a_j, der opfylder samtlige k ligninger. For at dette skal være tilfældet skal der dog yderligere gælde, at søjlerne i **X** er lineært uafhængige. Hvis systemet er underbestemt findes uendelig mange løsninger, svarende til et underrum af $\mathcal{R}^n$. Hvis søjlerne i **X** er lineært uafhængige er dimensionen af underrummet lig $n - k$. Hvis systemet er overbestemt findes der ingen løsninger **a** til dette. Givet et **a** kan fejlen, eller resudualen **r** beregnes ved $\mathbf{r} = \mathbf{y} - \mathbf{X}\mathbf{a}$. Et estimat af "løsningen" kan i stedet findes ved den vektor $\hat{\mathbf{a}}$, der minimerer normen af **r** (se kapitel 3). Nedenstående skitseret kort nogle metoder til "løsning" af kvadratiske og overbestemte ligningssystemer. For beskrivelse af metodernes beregningskompleksitet og numeriske egenskaber henvises til fagområdet *numerisk analyse*.

A.5 Løsning af kvadratiske lineære ligningssystemer

Antag at den reelle kvadratiske matricen $\mathbf{C}$ er regulær, dvs. at søjlerne i matricen er lineært uafhængige. Nedenstående skitserer en simpel og klassisk metode kaldet *Gauss-elimination.* Derefter skitseres nogle metoder, der baserer sig på dekomposition af matricen i "simplere" matricer. Læseren henvises til f.eks. [9] for egentlige algoritmebeskrivelser.

A.5.1 Gauss-elimination

Antag et kvadratisk lineært ligningssystem $\mathbf{Ca} = \mathbf{b}$ af dimension $n \times n$ givet. Den basale idé i Gauss eliminationsmetoden er at omforme sættet af ligninger således at der fremkommer et nyt ligningssystem $\mathbf{C}'\mathbf{a} = \mathbf{b}'$ hvor $\mathbf{C}'$ er en øvre trekantsmatrice, som illustreret nedenfor for $n = 4$.

$$\begin{bmatrix} c'_{11} & c'_{12} & c'_{13} & c'_{14} \\ 0 & c'_{22} & c'_{23} & c'_{24} \\ 0 & 0 & c'_{33} & c'_{34} \\ 0 & 0 & 0 & c'_{44} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} = \begin{bmatrix} b'_1 \\ b'_2 \\ b'_3 \\ b'_4 \end{bmatrix}$$

Dette ligningssystem kan løses ved baglængs substitution (eng. *back substitution*).

$$a_i = \frac{1}{c'_{ii}} \left[b'_i - \sum_{j=i+1}^{n} c'_{ij} a_j \right] \qquad \text{for } i = n, n-1, \ldots, 1$$

Omskrivningen af matricen $\mathbf{C}$ og vektoren $\mathbf{b}$ til matricen $\mathbf{C}'$ og vektoren $\mathbf{b}'$ foregår i $n-1$ skridt. I første skridt subtraheres skalerede versioner af den første ligning fra ligningerne under denne. Herved nulstilles alle elementer, bortset fra c_{11}, i første søjle af matricen. I det næste skridt gentages proceduren for den $(n-1) \times (n-1)$ delmatrice af $\mathbf{C}$ hvor første række og første søjle er udeladt.

I ovenstående procedure kan det ske at første element i første søjle af en delmatrice er lig 0. For at undgå division med 0 kan ligningerne ombyttes således at det numerisk største element i første søjle bliver det første element. Denne procedure kaldes *partiel pivotering.* En mere avanceret version er yderligere at ombytte søjlerne i delmatricen, således at det numerisk største element i hele delmatricen flyttes til første position i første søjle. Denne procedure kaldes *fuld pivotering.* Udgiften ved fuld pivotering er at positionen af de ubekendte a_i i vektoren $\mathbf{a}$ skal holdes ajour.

Gauss elimination med baglængs substitution har primært fordelen at være let at forstå. En ulempe ved metoden er at højresiden $\mathbf{b}$ skal være kendt. Gentagende løsninger med forskellig højreside kræver således at Gauss-eliminationen gentages for hver gang. Denne ulempe er ikke til stede ved de øvrige metoder beskrevet nedenfor.

A.5.2 LU-dekomposition og Cholesky dekomposition

En LU-dekomposition af en $n \times n$ reel matrice $\mathbf{A}$ er en opskrivning af denne som et produkt af en nedre trekantsmatrice $\mathbf{L}$ og en øvre trekantsmatrice $\mathbf{U}$. Herved kan det lineære ligningssystem opskrives:

$$\mathbf{Ax} \;=\; \mathbf{LUx} \;=\; \mathbf{L}(\mathbf{Ux}) \;=\; \mathbf{Lz} \;=\; \mathbf{y}$$

For $n = 4$ kan dekompositionen illustreres ved:

$$\begin{bmatrix} l_{11} & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 \\ l_{31} & l_{32} & l_{33} & 0 \\ l_{41} & l_{42} & l_{43} & l_{44} \end{bmatrix} \cdot \begin{bmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ 0 & u_{22} & u_{23} & u_{24} \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}$$

Antag at dekompositionen er foretaget, og lad $\mathbf{z}$ betegne (den endnu ukendte) vektor $\mathbf{Ux}$. Løsningen til det oprindelige system består nu i to etaper. Først løses:

$$\mathbf{Lz} \;=\; \mathbf{y}$$

Når vektoren $\mathbf{z}$ er fundet løses dernæst:

$$\mathbf{Ux} \;=\; \mathbf{z}$$

Fordelen ved LU-dekompositionen er at dekompositionen $\mathbf{A} = \mathbf{LU}$ er uafhængig af højresiden $\mathbf{y}$. En anden fordel ved LU-dekompositionen er at ovenstående to tridiagonale ligningssystemer er lette at løse vha. først en forlængs substitution:

$$z_i \;=\; \frac{1}{l_{ii}} \left[y_i - \sum_{j=1}^{i-1} l_{ij} z_j \right] \qquad \text{for } i = 1, 2, 3, \ldots, n$$

og derefter en baglængs substitution:

$$x_i \;=\; \frac{1}{u_{ii}} \left[z_i - \sum_{j=i+1}^{n} u_{ij} x_j \right] \qquad \text{for } i = n, n-1, n-2, \ldots, 1$$

Det vil føre for vidt her at beskrive metoder til beregning af de to matricer $\mathbf{L}$ og $\mathbf{U}$ ud fra $\mathbf{A}$.

Hvis matricen $\mathbf{A}$ vides at være symmetrisk og positiv definit er en *Cholesky dekomposition* ofte fordelagtig, fordi den beregningsmæssigt er hurtig (og numerisk stabil). Cholesky dekompositionen ligner LU-dekompositionen meget, idet matricen $\mathbf{U}$ i det aktuelle tilfælde kan vælges lig $\mathbf{L}^t$. Transformationen bliver således:

$$\mathbf{Ax} \;=\; \mathbf{LL}^t\mathbf{x} \;=\; \mathbf{y}$$

Når dekompositionen er foretaget kan ligningssystemet løses ved anvendelse af to baglængs substitutioner.

A.5.3 Egenværdi dekomposition

Lad $\mathbf{y} = \mathbf{A}\mathbf{x}$ være et kvadratisk ligningssystem, hvor $\mathbf{A}$ er reel og symmetrisk. Lad $\mathbf{P}$ være en matrice, bestående af egenvektorerne for $\mathbf{A}$, og lad λ_i være egenværdien hørende til den i'te egenvektor. Lad $\mathbf{D} = diag(\lambda_1, \lambda_2, \ldots, \lambda_n)$. Da gælder at:

$$\mathbf{x} \;=\; \mathbf{A}^{-1}\mathbf{y} \;=\; (\mathbf{P}\mathbf{D}^{-1}\mathbf{P}^{-1})\mathbf{y}$$

Da $\mathbf{P}^{-1} = \mathbf{P}^t$ ses løsningen af systemet, ud over bestemmelsen af egenvektorer og egenværdier, kun at bestå af matrice- og vektormultiplikationer. En fordel ved metoden er at egenværdierne, og hermed konditionstallet, er direkte til rådighed. Det er eksempelvis let at checke om matricen $\mathbf{A}$ er singulær. Det er ligeledes let at analysere i hvor høj grad de transformerede vektorer vil ligge i et underrum af $\mathcal{R}^n$. Vi skal ikke her diskutere metoder til beregning af egenvektorer og egenværdier.

A.6 Løsning af overbestemte lineære ligningssystemer

Der findes to grupper af metoder til løsning af overbestemte lineære ligningssystemer. I den første gruppe tages udgangspunkt i normalligningen (se nedenfor). Den anden gruppe af metoder udnytter det oprindelige sæt af ligninger. Lad $\mathbf{y} = \mathbf{A}\mathbf{x}$ være et overbestemt $k \times n$ ligningssystem, hvor rangen af $\mathbf{A}$ er lig n. *Normalligningen* fremkommer ved multiplikation med $\mathbf{A}^t$, dvs:

$$\mathbf{b} \;=\; \mathbf{A}^t\mathbf{y} \;=\; (\mathbf{A}^t\mathbf{A})\mathbf{x} \;=\; \mathbf{C}\mathbf{x}$$

Det fremkomne system er nu symmetrisk og kvadratisk ($\mathbf{C}$ er positiv definit), og de tidligere beskrevne metoder til løsning kan tages i anvendelse. Direkte opskrevet ses løsningen at være givet ved:

$$\mathbf{x} \;=\; (\mathbf{A}^t\mathbf{A})^{-1}\mathbf{A}^t\mathbf{y} \;=\; \mathbf{A}^{\#}\mathbf{y}$$

Matricen $\mathbf{A}^{\#}$ kaldes den *pseudoinverse* . Problemet ved brug af normalligningen er at konditionstallet for matricen $\mathbf{C}$ er kvadratet på konditionstallet for $\mathbf{A}$. Anderledes sagt, hvis rækkevektorene i $\mathbf{A}$ næsten er indeholdt i et underrum af $\mathcal{R}^n$, da vil metoden være numerisk ustabil. Da funktionen κ^2 vokser forholdsvist stærkt er metoden generelt problematisk ved løsning af overbestemte ligningssystemer.

A.6.1 Singulær værdi dekomposition

Singulær værdi dekomposition, SVD (Singular Value Decomposition) er “Rolls Royce” metoden for løsning af kvadratiske eller overbestemte ligningssystemer. Ved tillempning kan metoden også anvendes til analyse af underbestemte systemer. Metoden er beregningsmæssigt forholdsvis dyr, men giver en række fordele mht. fortolkning af transformationen. I en vis forstand er SVD en generalisation af egenværdi dekompositionen til ikke-kvadratiske matricer.

I det følgende antages om $m \times n$ matricen $\mathbf{A}$ at $m \geq n$. Iøvrigt er der ingen begrænsninger på matricen (den kan f.eks. være singulær). Det kan vises at $\mathbf{A}$ kan dekomponeres i et produkt af tre matricer $\mathbf{A} = \mathbf{UWV}^t$:

$$\left[\begin{array}{c} \\ \mathbf{A} \\ \\ \end{array} \right] = \left[\begin{array}{c} \\ \mathbf{U} \\ \\ \end{array} \right] \cdot \left[\begin{array}{cccc} w_1 & & & \\ & w_2 & & \\ & & \ddots & \\ & & & w_n \end{array} \right] \cdot \left[\begin{array}{c} \mathbf{V}^t \end{array} \right]$$

hvor $\mathbf{U}$ er en $m \times n$ søjle-ortogonal matrice, $\mathbf{W}$ er en $n \times n$ diagonalmatrice, og $\mathbf{V}$ er en $n \times n$ ortogonal matrice. Der gælder altså at:

$$\mathbf{U}^t\mathbf{U} \;=\; \mathbf{V}^t\mathbf{V} \;=\; \mathbf{V}\mathbf{V}^t \;=\; \mathbf{I}$$

I almindelighed vil $\mathbf{U}\mathbf{U}^t \neq \mathbf{U}^t\mathbf{U}$. Det følger at:

$$\mathbf{A}^{-1} \;=\; \mathbf{V} \cdot diag(\frac{1}{w_i}) \cdot \mathbf{U}^t$$

De *singulære værdier* w_i er enten lig 0 eller positive. Som ved en egenvædi dekomposition kan rangen af matricen $\mathbf{A}$ let findes som antallet af singulære værdier w_i, der er forskellige fra 0.

Hvis der findes singulære værdier w_i, der enten er lig 0, eller meget tæt på 0, er de tilsvarende elementer i diagonalmatricen $\mathbf{W}^{-1}$ enten udefinerede eller meget store. Konditionstallet κ for $\mathbf{A}$ er defineret ved den største singulære værdi divideret med den mindste singulære værdi. Hvis konditionstallet er meget stort (fordi $\mathbf{A}$ er næsten singulær) vil afrundingsfejl i beregningerne ofte resultere i en meget dårlig bestemt løsning. I dette tilfælde vil den (de) tilsvarende søjlevektorer i $\mathbf{U}$ of $\mathbf{V}$ være dårligt bestemt. Det er derfor tilrådeligt at ignorere den information som er indeholdt i det tilsvarende underrum. Dette gøres ved at definere $1/w_i$ til 0, hvis $w_i < T$, hvor T er en grænsevædi, der er bestemt ved maskinpræcisionen (f.eks. 10^{-6}). Løsningen til det kvadratiske eller overbestemte ligningssystem $\mathbf{Ax} = \mathbf{y}$ bliver således:

$$\mathbf{x} \;=\; \mathbf{V} \cdot [diag(r(w_i))] \cdot (\mathbf{U}^t\mathbf{y})$$

$$r(w_i) \;=\; \begin{cases} \frac{1}{w_i} & \text{hvis } w_i > T \\ 0 & \text{hvis } w_i \leq T \end{cases}$$

Som for de øvrige metoder skal vi ikke her angive algoritmer for beregning af dekompositionen.

A.7 Opgaver

1. Betragt matricen

$$\mathbf{M} \;=\; \left[\begin{array}{cc} 1 & 2 \\ 3 & 4 \end{array} \right]$$

Opstil det karakteristiske polynomium for matricen, og find egenværdierne. Bestem egenvektorerne (brug MATLAB) og afgør om disse er ortogonale. Bestem konditionstallet.

2. Betragt matricen

$$\mathbf{M} = \begin{bmatrix} 1 & 2 & 2 \\ 3 & 4 & 2 \\ 2 & 4 & 4 \end{bmatrix}$$

Bestem rangen af matricen. Opstil det karakteristiske polynomium for matricen, og find egenværdierne. Bestem egenvektorerne (brug MATLAB) for de positive egenværdier.

3. Betragt matricen

$$\mathbf{M} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \\ 3 & 1 & 4 \end{bmatrix}$$

Afgør om matricen har reelle egenværdier. Er matricen positivt definit? Hvad er konditionstallet?

4. Betragt matricen

$$\mathbf{M} = \begin{bmatrix} 1 & -1 & -1 \\ 1 & -1 & 0 \\ 1 & -1 & 1 \\ 1 & 0 & -1 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & -1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix}$$

Hvad er rangen af matricen? Bestem matricen $\mathbf{X} = \mathbf{M}^t\mathbf{M}$. Afgør om $\mathbf{X}$ har reelle egenværdier. Er $\mathbf{X}$ positivt definit? Hvad er konditionstallet? Beregn den pseudoinverse matrice.

5. Betragt ligningssystemet $\mathbf{Mx} = \mathbf{y}$, hvor:

$$\mathbf{M} = \begin{bmatrix} 0 & 2 & -2 \\ 3 & 4 & 2 \\ 2 & -4 & 3 \end{bmatrix}$$

og $\mathbf{y} = (1, 3, 7)^t$. Bestem rangen af $\mathbf{M}$. Omform ligningssystemet til et nyt system $\mathbf{M}'\mathbf{x} = \mathbf{y}'$, hvor $\mathbf{M}'$ er en øvre trekantsmatrice (brug partiel pivotering). Løs det ny system ved baglængs substitution.

Appendix B

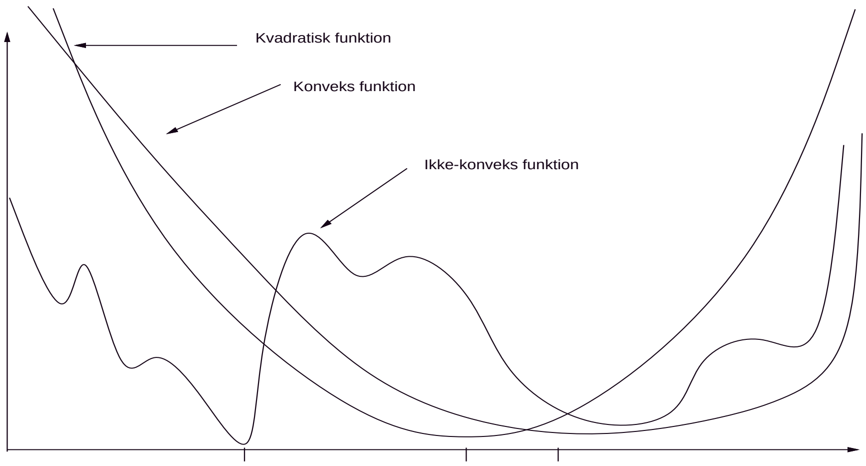
Optimeringsmetoder

Dette appendix er en kort introduktion til nogle ofte benyttede optimeringsmetoder. Sådanne metoder er bredt anvendt inden for mange grene af datalogien og udbudet af metoder er voldsomt. Det er således ikke muligt her at give en blot nogenlunde dækkende beskrivelse af dette felt. Eksempelvis vil metoder til diskret optimering ikke blive omtalt. I kontinuert optimering søges den værdi, der maksimerer eller minimerer en funktion. Valget af optimeringsmetode afhænger i høj grad af hvorledes funktionen ser ud. Et formål med afsnittet er klassificere funktionerne således at et hensigtsmæssigt valg af optimeringsmetode bliver lettere. Et andet formål er at afrunde de diskussioner om estimationsmetoder, som vi efterlod i tidligere afsnit af noterne. Eksempelvis angiver et Bayes estimat hvad der skal optimeres, men beskriver ikke hvordan den optimale løsning skal findes.

Fra gymnasiestoffet vides at en funktion $f(x)$ kan antage et ekstremum i punktet x_0 hvis den afledede $f'(x_0) = 0$. Denne viden vil være nyttig ved bestemmelse af et maksimum likelihood eller et maksimum a posteriori estimat af en parameter. Et lille problem er, at kriteriet er opfyldt for minima såvel som for maksima (samt for saddelpunkter). Det er dog ofte let af afgøre (f.eks. ved undersøgelse af f'') om der er tale om et minimum eller et maksimum. Et større problem er at funktionen f kan indeholde flere lokale ekstrema. Ved optimeringsopgaver karakteriserer man ofte funktionen efter hvorvidt den er *kvadratisk*, *konveks*, eller *ikke-konveks*. I afsnittet om lineær regression så vi at miminering af et udtryk af formen:

$$f(\boldsymbol{\theta}) = \sum_i (y_i - \mathbf{x}_i^t \boldsymbol{\theta})^2 \tag{B.1}$$

resulterede i løsningen af et overbestemt lineært ligningssystem. Årsagen til dette er, at udtrykket er en sum af led, der hver er kvadratisk i den ubekendte $\boldsymbol{\theta}$. Denne yderst behagelige situation tillader anvendelse af hurtige (lineære) metoder. Mange fuktioner (f.eks. fordelingsfunktioner) er ikke kvadratiske. Ved optimering af sådanne funktioner kan vi ikke regne med at den lineære algebra kan hjælpe os med at finde den entydige løsning. Det kan være vanskeligt at afgøre om en ikke-kvadratisk funktion har ét eller flere ekstrema. Man skelner i stedet mellem konvekse og ikke-konvekse funktioner. For en konveks funktion $f(\mathbf{x})$ gælder at den rette linie mellem to vilkårlige punkter $(\mathbf{x}_1, f(\mathbf{x}_1))$



Figur B.1: Illustration af en kvadratisk, en konveks, og en ikke-konveks funktion.

og $(\mathbf{x}_2, f(\mathbf{x}_2))$ ligger helt på den ene side af (typisk over) funktionen (området over funktionen er konvekst). Alle kvadratiske funktioner er således konvekse, og mange konvekse funktioner har kun ét ekstremum. Derimod kan en funktion med flere lokale ekstrema ikke være konveks. Dette er årsagen til at konveksiteten af en funktion ofte er afgørende ved optimering af funktionen.

Gradient descent-metoden, der beskrives nærmere nedenfor, er en metode til bestemmelse af ekstrema for konvekse funktioner. Som beskrevet ovenfor er kravet om konveksitet i virkeligheden for stærkt. Metoden vil således kunne fungere tilfredsstillende på mange ikke-konvekse funktioner. Gradient descent metoden er formodentlig en af de mest benyttede optimeringsmetoder overhovedet. Dette skyldes at mange praktiske problemer kan løses ved at opstille et konvekst udtryk (en straf-funktion) for kvaliteten af løsningen.

I mange mere komplicerede situationer er den funktion, der ønskes minimeret ikke konveks. Dette er eksempelvis tilfældet med stereo-eksemplet i Appendix C. Her er funktionalen afhængig af billederne I_L og I_R. Hvis man forestiller sig disse som sinus-bølger, vil kravet om approsimativt ens intensitet i korresponderende punkter være lige godt opfyldt for et antal positioner svarende til bølgelængden τ. Alle funktioner af formen $d_x = c + n\tau$, hvor c er en konstant, og n et vilkårligt heltal, vil derfor være lige gode løsninger. I andre situationer hvor flere krav ønskes opfyldt samtidigt, bliver funktionen ofte ikke-konveks. I så tilfælde vil det lokal minimum, som en gradient descent metode vil angive, ikke nødvendigvis være det globale minimum. Hvis funktionalen indeholder deltafunktioner er optimeringsproblemet særlig ubehageligt, fordi de lokale minima kan ligge vilkårligt tæt. Der findes ingen beregningsmetoder, der med sikkerhed kan finde det globale minimum af enhver ikke-konveks funktion inden for endelig tid. Derimod findes gode metoder, hvormed en approksimation til det globale minimum kan findes inden for acceptabel (men ofte lang) tid.

I afsnit 2 og 3 skitseres to klasser af optimeringsmetoder. Den ene er baseret på stokastisk optimering, den anden på gentagende substitution af funktionen, med en simplere. I begge tilfælde ønsker vi at bestemme minimum af en funtion f af en vektorvariabel $\mathbf{x}$. Ønskes maksimum for $f(\mathbf{x})$ fundet skal vi ækvivalent finde minimum for $-f(\mathbf{x})$. De nærmere krav til funktionen (f.eks. differentiabilitet e.lig.) omtales undervejs.

B.1 Gradient descent

Gradient descent metoden kan illustreres ved at betragte den funktion, der ønskes minimeret (ofte kaldt en *straffunktion* eller en *energifunktion*) som en terrænmodel, hvor vi ønsker at finde den dybeste dal. Ved at lade en bold rulle (under påvirkning af tyngdekraften) fra en vilkårlig position vil den ende i et lokalt minimum. Hvis terrænet er konvekst vil dette kun indeholde ét minimum, der derfor er globalt. Den banekurve, som bolden vil følge vil (idet vi ignorerer at boldens har en inerti og derfor opnår en kinetiske energi) være karateriseret ved niveaukurverne for terrænet. På ethvert tidspunkt vil bolden tage en retning, der er vinkelret på niveaukurven gennem punktet. Der findes to sådanne retninger: den ene peger nedad, den anden opad. Retningen opad er fastlagt ved gradienten til terrænmodellen. Retningen nedad er bestemt ved den negative gradient. Når gradient descent metoden skal benyttes til minimering af funktionen $f(\mathbf{x})$ er det derfor en forudsætning at gradienten ∇f kan beregnes.

Antag at dimensionen af $\mathbf{x}$ er n og lad:

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots \frac{\partial f}{\partial x_n}\right) \tag{B.2}$$

Lad $\mathbf{x}_0$ være et vilkårligt startpunkt i det n-dimensionale parameterrum, og lad k være en reel konstant der angiver *skridtlængden*. Lad ϵ angive præcisionen hvormed minimumspunktet for funktionen ønskes fundet. En simpel gradient descent metode, der går under navnet *steepest descent* kan skitseres ved:

For $\mathbf{x} = \mathbf{x}_0$ indtil $||\nabla f(\mathbf{x})|| < \epsilon$:

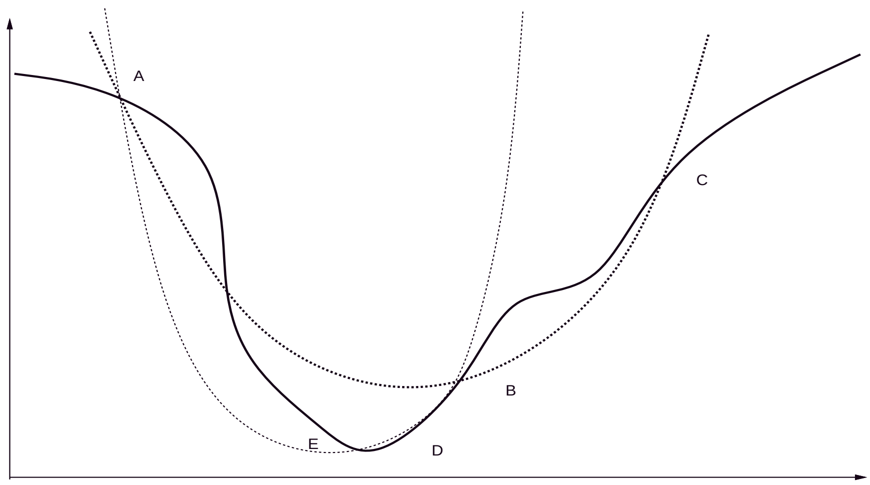
1. Lad $\mathbf{v} = \frac{\nabla f(\mathbf{x})}{||\nabla f(\mathbf{x})||}$.
2. Lad $\mathbf{x} := \mathbf{x} - k\,\mathbf{v}$

Metoden består således i skridtvis at flytte punktet $\mathbf{x}$ i retning modsat gradienten. Skridtlængden er i ovestående skitse valgt konstant lig k. Bemærk at banekurven kun vil approksimere boldens bane for $k \to 0$. Da stopkriteriet ideelt set er, at længden af gradienten er lig 0 ses, at hvis k er valgt uheldigt, så vil dette punkt aldrig blive nået. Bolden vil hele tiden rulle for langt. Det er derfor nødvendigt at mindske skridtlængden k. Hvis k vælges meget lille vil gradienten blive udregnet mange gange, hvilket vil tage lang tid. Hvis k vælges stor vil vi ikke være i stand til at lokalisere minimumspositionen. Generelt kan vi ikke bestemme denne med en præcision, der er større end k. Det er derfor nødvendigt at modificere ovenstående skitse således at værdien af $f(\mathbf{x})$ beregnes løbende. Denne værdi skal gerne aftage. Hvis den vokser, er det nødvendigt at mindske

k. Bestemmelse af skridtlængden k, hvorledes den skal reduceres, samt præcisionsangivelsen ϵ er helt afhængig af den aktuelle anvendelse, og kan ikke fastlægges uden et godt kendskab til funktionens opførsel.

Der findes mange varianter og forbedringer af ovenstående simple skitse. Vi skal her kun diskutere én sådan forbedring. Antag at funktionen f indeholder en smal dal, der snor sig nedad mod det globale minimum. Antag videre at $\mathbf{x}_0$ angiver et punkt på dalens side. Banekurven for postitionerne af $\mathbf{x}$ vil da komme til at ligne et sik-sak mønster, der følger dalen. I stedet for at følge bunden af dalen direkte ned mod det globale minimum, vil ruten løbe frem og tilbage næsten på tværs af dalen (men dog nedad). Der findes flere måder at løse problemet på. I én metode erstattes retningsvektoren bestemt af gradienten med middelværdien af den aktuelle og den forrige retningsvektor. Dette vil udglatte en sik-sak rute. En anden metode er at lokalisere bunden af dalen langs det endimensionale snit af funktionen f givet ved punktet $\mathbf{x}$ og gradientretningen gennem dette punkt. Når punktet i bunden af dalen er fundet genberegnes gradienten og processen gentages. Bunden af dalen kan ofte findes med acceptabel præcision ved iterativ approksimation af det endimensionale snit med et anden ordens polynomie. Givet tre positioner a, b, og c langs snittet f_s vil minimumspunktet y for det approksimerende anden grads polynomie være:

$$y \;=\; b \;+\; \frac{1}{2}\,\frac{(b-a)^2[f_s(b)-f_s(c)] \;-\; (b-c)^2[f_s(b)-f_s(a)]}{(b-a)[f_s(b)-f_s(c)] \;-\; (b-c)[f_s(b)-f_s(a)]} \tag{B.3}$$

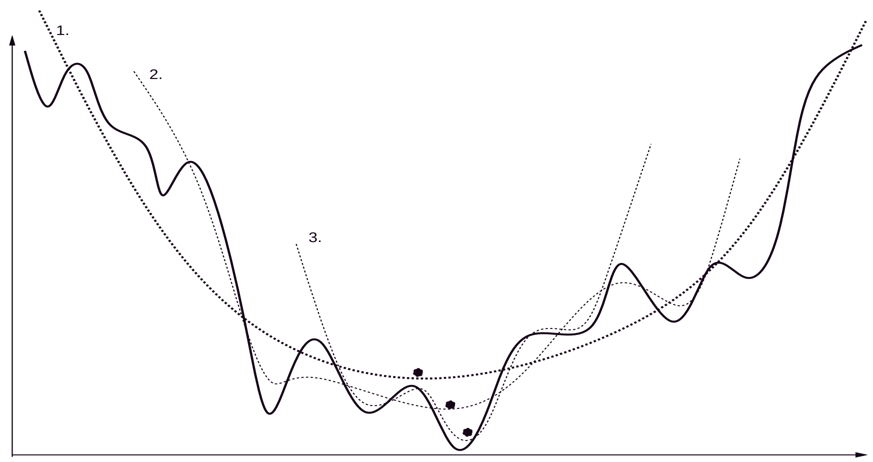


Figur B.2: Illustration af en simpel metode til minimering af en endimensional funktion. Først bestemmes punktet D, der minimerer værdien af parablen gennem A, B og C. I andet trin findes minimum E af parablen gennen A, D og B.

Efter bestemmelse af punkterne a, b, og c samt beregning af y, erstattes enten a med b (hvis $a < b < y < c$) eller c med b (hvis $a < y < b < c$). Desuden erstattes b med y. Et nyt fit foretages, og processen gentages indtil det endimensionale minimum er fundet med tilstrækkelig præcision.

B.2 Variationsmetoder

Hvis funktionen har flere lokale minima, vil enhvert lokalt minimum fundet af en gradient descent metode ikke nødvendigvis være globalt. En metode til valg af lokalt minimum er, at erstatte den oprindelige funktion med en konveks funktion, der i grove træk ligner den oprindelige funktion. Minimumspunktet for den konvekse funktion kan findes ved en gradient descent metode, og vil angive et groft skøn over positionen af det mindste lokale minima for den oprindelige funktion. Ved nu at variere den approksimerende funktion gradvist mod den oprindelige, og løbende spore minimumspunktet kan det globale minimum ofte findes. Denne variationsteknik (eng. Method of continuation), er illustreret i nedenstående figur.



Figur B.3: Illustration af en variationsmetode. Den optrukne kurve viser den oprindelige funktion. De tre stiplede kurver viser gradvist mindre udglattede versioner af den oprindelige funktion. De sorte kugler illustrere de fundne minima.

Metoden sikrer på ingen måde at det korrekte minimum findes. Dette afhænger meget af funktionens udseende, af variationshastigheden, samt hvorledes substitut-funktionen bestemmes. Det er let at konstruere eksempler hvor metoden vil fejle. I mange praktiske eksempler vil metoden imidlertid give det korrekte resultat. Fordelen ved metoden er, at selv om der skal foretages et (større) antal gradient descent operationer, vil beregningstiden ofte være betragteligt mindre end for metoder, der benytter stokastisk optimering.

Det største problem ved anvendelse af variationsmetoder er at bestemme sekvensen af approksimationer til den oprindelige funktion. Der findes, forfatteren bekendt, ingen generelle retningslinier for dette. I enkelte tilfælde er velfungerende metoder dog kendte. Den interesserede læser kan finde inspiration bl.a. i [2].

B.3 Stokastiske optimering

Metoder til *Stokastisk optimering* hører formodentlig til de mest generelle, og de mest ressourcekrævende, optimeringsmetoder. Metoderne er baseret på stokastisk stikprøvetagning. Hvis løsningsrummet er endeligt (men stort) kan stokastisk sampling benyttes til at reducere mængden af beregninger. Dette sker på bekostning af en risiko for ikke at finde den optimale løsning. Nedenstående gennemgås denne metode med udgangspunkt i mindste medianers metode til lineær regression (se Appendix C). Derefter beskrives en generel stokastisk optimeringsmetode, kaldet simuleret udglødning, der i de senere år har påkaldt sig stor opmærksomhed.

B.3.1 Stokastisk sampling

I mindste medianers metode til lineær regression skal medianværdien af residualvektoren beregnes for samtlige kombinationer hvormed vi kan udtage n ligninger blandt de k observationer (her antages at der er givet k ligninger i n ubekendte, $k \gg n$). Hvis antallet n af modelvariable er blot moderat stort vil antallet af gange, som medianværdien skal udregnes, være astronomisk. Ved stokastisk sampling nøjes man med kun at udtage en lille delmængde af de mange mulige kombinationer. Dette sker, naturligvis, på bekostning af sikkerheden af estimationen. Hvis formålet er at sortere outliers fra, er det imidlertid tilstrækkeligt hvis alle udtagne n ligninger i blot én stikprøve er inliers. Regressionen foretaget på denne stikprøve vil være korrekt (dog næppe nøjagtig). I det følgende skal vi opstille et udtryk for sandsynligheden for, at der blandt m stikprøver findes mindst én, der ikke indeholder outliers.

Antag er $k \gg n$, og at andelen af outliers blandt de k observationer er ϵ. Ved tilfældigt valg af n ligninger fremkommer en potentiel løsning som vi karakteriserer ved medianen af de til fittet hørende residualværdier. Sandsynligheden for at trække n inliers er lig $(1-\epsilon)^n$. Sandsynligheden for, at der blandt de n observationer findes mindst én outlier er derfor $1-(1-\epsilon)^n$. Lad os foretage m stikprøver. Sandsynligheden for, at der blandt disse findes mindst én prøve, der udelukkende indeholder inliers er:

$$1 \; - \; [1 \; - \; (1-\epsilon)^n]^m \tag{B.4}$$

Ved at kræve at denne sandsynlighed er stor (tæt på 1, f.eks. 0.99) kan vi bestemme hvor mange gange vi skal udtage de n observationer. Lad v angive den krævede sandsynlighed for succes. Da får vi at:

$$m \; > \; \frac{\log(1 \; - \; v)}{\log(1 \; - \; [1-\epsilon]^n)} \tag{B.5}$$

Eksempel
Antag at der er givet 100 ligninger hver med n ubekendte, og at vi vil kræve en sikkerhed i estimationen på 99 %. Nedenstående skema viser for $n = 2, 3, \ldots 10$, og for $\epsilon = 0.1, 0, 2, 0.5$ antallet af gange residualvektoren skal beregnes ved stokastisk sampling. Desuden er angivet det tilsvarende tal for udtømmende søgning. Som det fremgår tilbyder stokastisk

sampling her en voldsom hastighedsforøgelse på bekostning af en lille sandsynlighed for at begå fejl.

n	Udtømmende søgning	10% outliers	20% outliers	50% outliers
2	4.950	3	5	16
3	161.700	4	6	35
4	3.921.255	5	9	72
5	75.287.520	6	12	145
6	1.192.052.400	6	16	293
7	16.007.560.800	7	20	587
8	186.087.894.300	9	25	1177
9	1.902.231.808.400	10	32	2356
10	17.310.309.456.440	11	41	4714

Eksempel slut

B.3.2 Simuleret afkøling

Simuleret afkøling (også kaldt simuleret udglødning, simuleret nedkøling mv.) er en metode, der er inspireret af krystaldannelse. Energien af en kold krystal er defineret ved summen over alle molekyler i krystallen af deres potentielle energi. Ved opvarmning af krystallen tilføres termisk energi. Afhængig af mængden af termisk energi kan molekylerne bevæge sig mere eller mindre frit. Efter opvarmningen afkøles krystallen. Herved fjernes termisk energi. Hvis afkøling sker hurtigt vil krystallen ofte have uregelmæssigheder. Hvis krystallen afkøles meget langsomt er det muligt at krystallen bliver helt perfekt (molekylerne er arrangeret i et perfekt gitterværk). Energien for den perfekte krystal vil være minimal. I simuleret afkøling forsøges denne proces efterlignet. Dette kræver at der til enhver løsning (molekylekonfiguration) kan associeres et energimål E.

For en given temperaturværdi T er det *termale equilibrium* karakteriseret ved (en konstant) fordeling af energiværdi. For stor T er fordelingen meget bred, for lille T er fordelingen koncentreret om den aktuelle konfiguration. Fordelingen af energiværdi ændrer sig således under af afkølingen. Hvis temperaturen holdes konstant opstår et termalt equilibrium. Fordelingen af energi E ved temperaturen T kan vises at være bestemt ved:

$$p_T(E) \;=\; \frac{1}{Z(T)} exp(-\frac{E}{kT}) \tag{B.6}$$

hvor $Z(T)$ er en normaliseringsfaktor (eng. partition function), og hvor k er en konstant (kaldet Boltzmann konstanten). Fordelingen selv kaldes ofte for *Boltzmann fordelingen.*

Inden for et kort tidsinterval dt bevæger molekylerne sig mere eller mindre tilfældigt med en hastighed, der er afhængig af temperaturen. Givet en molekylekonfiguration K_t til tiden t er det derfor rimeligt at antage at mængden af opnåelige konfigurationer til tiden $t + dt$ er parametriseret af temperaturen T. Vi vil kalde denne mængde for en *naboomegn* og betegne den ved $\mathcal{N}(K_t, T)$.

Metropolis algoritmen er en meget kendt og anvendt metode til at simulere udviklingen mod et termalt equilibrium. En skitse af algoritmen er angivet nedenstående. Her er energien E parametriseret af den konfiguration $\boldsymbol{\theta}$ som vi ønsker at bestemme.

1. Lad $\boldsymbol{\theta}$ være en konfiguration med energi $E(\boldsymbol{\theta})$.
2. Gentag:
   - Lad $\boldsymbol{\theta}_{ny}$ være en ny konfiguration valgt tilfældigt blant alle nabokonfigurationer i $\mathcal{N}(K_t, T)$. Lad $\boldsymbol{\theta}_{ny}$ have energien $E(\boldsymbol{\theta}_{ny})$, og lad $\Delta = E(\boldsymbol{\theta}_{ny}) - E(\boldsymbol{\theta})$.
   - Hvis $\Delta \leq 0$ eller hvis $exp(-\frac{\Delta}{kT}) > random([0,1[)$, så flyt konfiguration, dvs. lad $\boldsymbol{\theta} := \boldsymbol{\theta}_{ny}$.
3. Indtil equilibrium er fundet.

Så længe en ny konfiguration med lavere energi findes, flyttes den aktuelle konfiguration dertil. Hvis den ny konfiguration har større energi flyttes kun med en vis sandsynlighed, beskrevet ved Boltzmann fordelingen. Dette muliggør, at optimeringen kan undslippe lokale minima. Efter tilstrækkeligt mange iterationer kan det vises, at en ligevægtstilstand indtræder. I denne flyttes lige mange gange til en konfiguration med højere energi, som til én med lavere energi. Det kan være vanskeligt at specificere klart hvornår ekvilibrium indtræffer. I mange algoritmer benyttes et fast (stort) antal iterationer.

En algoritme til simuleret afkøling kan nu konstrueres ved i trin langsomt at sænke temperaturen, og først fortsætte på næste trin når ekvilibrium er fundet:

1. Lad $T = T_{start}$.
2. Gentag:
   - Benyt Metropolis algoritme for temperatur lig T.
   - Sænk temperaturen $T := g(T)$.
3. Indtil stopkriterium nået.

Temperaturen sænkes ofte med en konstant faktor, dvs. $g(T) = \alpha T$, hvor α er en konstant tæt på 1 (f.eks. 0.9). Stopkriteriet består typisk af en kombination af et loft over antallet af iterationer, og et krav om en gradvis formindskelse af energivineauet.

Der er, som det fremgår, intet krav til energifunktionen E, bortset fra, at denne skal være ikke-negativ. Metoder er derfor meget velegnet i situationer hvor E ikke er konveks. Metoden er meget generel, men også ekstremt beregningskrævende. Et problem ved anvendelse af metoden at definere naboområderne $\mathcal{N}$. Ved optimering over et kontinuert domæne er det naturligt at bestemme $\mathcal{N}$ som en kugle (med centrum i den aktuelle konfiguration) og med radius bestemt ved en aftagende funktion $r(T)$ af temperaturen. Denne tilgang er imidlertid farlig, fordi visse komponenter af parametervektoren ofte varierer meget mere end andre. En anden oplagt tilgang er løbende at estimere kovariansmatricen

$\mathbf{C}$ for mængden af de seneste succesfulde konfigurationer. Disse kunne f.eks. være karakteriseret ved en lavere energi end deres umiddelbart foregående konfigurationer. Dernæst kunne mængden af nabokonfigurationer bestemmes som mængden af $\boldsymbol{\theta}_{ny}$ hvorom:

$$[\boldsymbol{\theta}_{ny} - \boldsymbol{\theta}]^t \mathbf{C}^{-1} [\boldsymbol{\theta}_{ny} - \boldsymbol{\theta}] \leq r(T)$$

hvor $r(T)$ er en (f.eks. esponentielt) aftagende funktion af T. Ovenstående forslag har imidlertid en stor beregningsmæssig ulempe idet kovariansmatricen skal estimeres og inverteres løbende under optimeringen. Vi skal ikke her diskutere de mange varianter af optimering ved simuleret udglødning yderligere. Intersesserede læsere henvises til f.eks. [7].

B.3.3 Genetiske algoritmer

Genetiske algoritmer er en gruppe af metoder der tager udgangspunkt i at simulere genudviklingen i en population. Inspirationen er Darwins evolutionslære. Kravet til anvendelse af metoden er, at parametrene kan repræsenteres som gener (mere om dette senere), og at der til ethvert individ (dvs. værdi af parametervektoren) kan knyttes et (positivt) godhedsmål. Hensigten med metoden er at maksimere summen af godhed over hele populationen.

En simpel genetisk algoritme fungerer ved iterativt at udvælge to individer fra populationen, lade disse få to børn, og derefter dræbe to individer fra populationen. Størelsen af populationen holdes altså konstant. Hvert individ udvælges til forælder med en sandsynlighed, der svarer til individets forholdsmæssige godhed i populationen. Hensigten er, at simulere et ønske om, at gode forældre får gode børn. De der dræbes kan udvælges tilfældigt, eller som de der er mindst gode e.lign. Børnene dannes som individer, hvis gener fås ved overkrydsning af forældrenes gener. Hertil kommer eventuelt en mutation. Denne proces er illustereret i nedenstående eksempel.

Eksempel
Lad populationen bestå af 5 individer, hver specificeret af 8 binære gener. Lad godheden af hvert individ være givet som anført i nedenstående skema:

individ	population								godhed	sandsynlighed
1	1	0	1	1	0	0	1	0	4	0.16
2	1	1	0	1	1	0	1	0	7	0.28
3	0	1	1	1	0	1	0	1	3	0.12
4	1	1	0	1	0	0	1	0	6	0.24
5	1	1	1	0	1	0	0	1	5	0.20
									25	1.00

Efter de anførte sandsynligheder trækkes nu tilfældigt individ nummer 2 og nummer 5 som forældre. Genkrydsningen foretages ved at først at bestemme hvilke gener der skal krydses. Der kan tænkes mange måder at gøre dette på. Ved ét-punkts krydsover bestemmes først en genposition t. Det ene barn arver de første t gener fra moderen og de resterende gener fra faderen. Det andet barn arver de de første t gener fra faderen og de resterende gener fra moderen. Her skal vi benytte en to-punkts krydsover. Antag at gen nummer 3-5 overkrydses. De to børn dannes nu som nye individer hvor det ene er identisk med med moderen, bortset fra at gen 3-5 er arvet fra faderen. Det andet barn er identisk med faderen, bortset fra at gen 3-5 er arvet fra moderen. Dette er illustreret nedenstående hvor de to børn er givet individnumrene 6 og 7.

individ	forældre og børn								godhed
2	1	1	0	1	1	0	1	0	7
5	1	1	1	0	1	0	0	1	5
6	1	1	1	0	1	0	1	0	6
7	1	1	0	1	1	0	0	1	6

Tilfældigt dræbes nu individerne 4 og 1. Den ny population bliver derefter:

individ	population								godhed	sandsynlighed
2	1	1	0	1	1	0	1	0	7	0.26
3	0	1	1	1	0	1	0	1	3	0.11
5	1	1	1	0	1	0	0	1	5	0.19
6	1	1	1	0	1	0	1	0	6	0.22
7	1	1	0	1	1	0	0	1	6	0.22
									27	1.00

Eksemplet illustrerer at godheden af populationen efter generationsskiftet er forbedret fra 25 til 27.

Eksempel slut

Ovenstående simple skitse kan modificeres i det uendelige. Eksempelvis kan individer tilskrives en alder, og dø efter et tidsrum etc. Sådanne modifikationer har ofte ingen umiddelbar relevans ved brug i bundne minimeringsopgaver. Derimod kan det vises at mutation er en vigtig måde at forhindre indavl. Problemet er, at relativt dårlige gener hurtigt sorteres fra, hvorved genpuljen bliver lille. I termer af optimering er denne fanget i et lokalt minimum. Selv om populationen er forholdsvis stor, skaber mutation, i lighed med det tilfældige element i simuleret afkøling, den nødvendige evne til at undslippe lokale ekstrema.

Et væsenligt problem ved anvendelse af genetiske algoritmer er at repræsentere parametrene i optimeringsproblemet som gener. I ovenstående eksempel antog hvert gen en binær værdi. I praksis ønsker vi ofte at optimere over et vektorrum af reelle værdier.

Hvis disse ikke er arrangeret på en måde sådan at krydsningen af individer med af gode gener generelt skulle lede til gode individer, da kan metoden ikke forventes at konvergere hurtigere end ved tilfældig permutation af generne.

B.4 Opgaver

1. Lad $f(x) = log(1+x^2)$. Vis at funktionen ikke er konveks. Vis at funktionen kun har ét minimum og er monotont voksende hhv. aftagende for $x > 0$ hhv. $x < 0$. Benyt *Stepest descent*-metoden til at finde det globale minimum idet startpunktet x_0 vælges til 5.0, $\epsilon = 0.001$ og skridtlængden bestemmes ved kvadratroden af gradientstørelsen. Skriv et program (eller benyt en lommeregner) til minimeringen. Beskriv uformelt din erfaring med metoden i det konkrete tilfælde.

2. Skriv et program, der implementerer mindste medianers metode til lineær regression under brug af stokastisk sampling. Antag at dimensionen af $\mathbf{x}$-vektoren er 1, dvs. at problemet er karakteriseret ved k ligninger af formen $y_i = a_0 + a_1 x$. Evaluér eksperimentelt træfsikkerheden af algoritmen (ved korrekt angivelse af ϵ).

3. Diskutér sammenhængen mellem temperaturen T og naboomegnene $\mathcal{N}$ for den simulerede afkøling.

4. Implementér, ved simuleret afkøling, en metode til estimation af parametrene a_0 og a_1 svarende til beskrivelsen i opgave 2. Hvordan er energifunktionen defineret? Hvordan bestemmes naboområder? Hvorledes fungerer metoden i praksis, og hvor hurtig er den?

5. Lad $f_0(x)$ være en kontinuert funktion, der er implicit givet ved n stikprøver (x_i, y_i), $i = 1 \ldots n$. Hvor mange funktioner g stemmer overens med f i de givne stikprøver? Glatheden af en differentiabel funktion g defineret på intervallet [a:b] kan angives ved $\int_a^b [g'(x)]^2 dx$. Angiv et udtryk (en energifuktion), der til hver differentiabel funktion g udtrykker at denne både skal stemmme godt overens med de givne stikprøver, og være mest mulig glat. Formulér den tilsvarende energifunktion hvis g er diskret (Benyt en differensapproksimation til g'). Er udtryket kvadratisk, konvekst?

Appendix C

Mere Statistisk Dataanalyse

I dette afsnit uddybes nogle emner fra statistisk dataanalyse, som kun er perifert berørt i noteafsnittet.

C.1 Basal Sandsynlighedsregning

I det følgende skal vi betragte reelle stokastiske (tilfældige) variable. Et *sandsynlighedsfelt* består af to komponenter, en mængde Ω og en funktion p. Mængden Ω, kaldes *udfaldsrummet*, og elementerne i Ω kaldes udfald. En delmængde af Ω kaldes en *hændelse*. Lad $\mathcal{S}$ være mængden af mulige hændelser ($\mathcal{S}$ bør være en Borel-mængde, hvorom der kræves lidt flere krav, der ikke skal omtales her). p er en funktion, der til hver hændelse A lader svare et reelt tal $p(A)$, kaldet *sandsynligheden* for hændelsen A. Om funktionen $p : \mathcal{S} \rightarrow \mathcal{R}$ gælder der (idet de ikke nævnte forudsætninger udnyttes) at:

1. $0 \leq p(A) \leq 1 \quad \forall A \in \mathcal{S}$
2. $p(\Omega) = 1$
3. $p(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} p(A_i)$ hvis der gælder at $\forall i \neq j : A_i \cap A_j = \emptyset$

Det vises let at $p(\emptyset) = 0$, og at $p(\Omega \backslash A) = 1 - p(A)$.

Et sandsynlighedsfelt kaldes *diskret* hvis Ω er numerabel, og kaldes et *Laplacefelt* hvis $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ er endelig og hvis alle hændelser har samme sandsynlighed, dvs. $p(\omega_i) = 1/n$.

Lad A og $B \neq \emptyset$ være to hændelser. Den *betingede sandsynlighed* $p(A|B)$ er defineret ved:

$$p(A|B) \;=\; \frac{p(A \,\cap\, B)}{p(B)} \tag{C.1}$$

De to hændelser siges at være *uafhængige* hvis der gælder at:

$$p(A|B) \;=\; p(A) \tag{C.2}$$

Generelt gælder at:

$$p(A|B) \;=\; \frac{p(A \,\cap\, B)}{p(B)} \;=\; \frac{p(B \,\cap\, A)}{p(A)} \frac{p(A)}{p(B)} \;=\; \frac{p(B|A)\, p(A)}{p(B)} \tag{C.3}$$

Denne identitet kaldes *Bayes regel.*

Eksempel
Lad $\Omega = \{0,1,2,3,4,5,6,7,8,9\}$ være mængden af cifre, og lad sandsynligheden for en hændelse A være lig kardinaliteten af A divideret med kardinaliteten af Ω, dvs. $p(A) = |A|/|\Omega| = |A|/10$. Lad $A = \{1,4,5,8\}$ og lad $B = \{1,5,9\}$. Da er $p(A) = 4/10$ og $p(B) = 3/10$. Den betingede sandsynlighed for at et udfald tilhører hændelsen A givet at vi ved at det tilhører hændelsen B er:

$$p(A|B) \;=\; \frac{p(A\cap B)}{p(B)} \;=\; \frac{p(1,5)}{p(1,5,9)} \;=\; \frac{2/10}{3/10} \;=\; \frac{2}{3}$$

Tilsvarende fås:

$$p(B|A) \;=\; \frac{p(1,5)}{p(1,4,5,8)} \;=\; \frac{1}{2} \;=\; \frac{p(A|B)p(B)}{p(A)} \;=\; \frac{(2/3)(3/10)}{4/10}$$

Eksempel slut

Lad (Ω, p) være et sandsynlighedsfelt. En *stokastisk variabel* x er en reel funktion defineret på Ω. Ved *fordelingen* af en stokastisk variabel $x : \Omega \rightarrow \mathcal{R}$, forstås den funktion p_x, der til en delmængde Π af $\mathcal{R}$ bestemmer sandsynligheden for at $x \in \Pi$, dvs.:

$$p_x(\Pi) \;=\; p(\{\omega | x(\omega) \in \Pi\}) \;=\; p(x^{-1}(\Pi)) \tag{C.4}$$

Fordelingsfunktionen F for en stokastisk variabel $x : \Omega \rightarrow \mathcal{R}$ er en funktion $F : \mathcal{R} \rightarrow [0,1]$ defineret ved:

$$F(t) \;=\; p(x \leq t) \;=\; p_x(]-\infty, t]) \tag{C.5}$$

Fordelingsfunktionen er lettere at anvende i praksis fordi den ikke kræver kendskab til $p(\Pi)$ for samtlige delmængder Π af $\mathcal{R}$. To stokastiske variable med samme fordeling har samme fordelingsfunktion. Om fordelingsfunktionen F gælder at:

1. $\lim_{x\rightarrow -\infty} F(x) = 0$
2. $\lim_{x\rightarrow \infty} F(x) = 1$
3. F er voksende
4. F er kontinuert fra højre
5. $\lim_{x\rightarrow t+} F(x) \;-\; \lim_{x\rightarrow t-} F(x) \;=\; p(x = t)$

En stokastisk variabel x kaldes kontinuert hvis der $\forall t \in \mathcal{R}$ gælder at $p(x = t) = 0$. I dette tilfælde siges x at have en kontinuert fordeling. *Tæthedsfunktionen* for en kontinuert stokastisk variabel x med fordeling F er en funktion f givet ved:

$$f(x) = F'(x) \quad \forall x \in \mathcal{R} \tag{C.6}$$

Om tæthendsfunktionen gælder oplagt at $\int_a^b f(x)dx \;=\; F(b) - F(a)$, og at $\int_{-\infty}^{\infty} f(x)dx \;=\; 1$, og at $F(x) = \int_{-\infty}^{x} f(t)dt$.

C.1.1 Sammenligning af fordelinger

Sammenligning af fordelinger er en særdeles nyttig operation i mange metoder til mønstergenkendelse. I mange klassifikationsmetode (se Kapitel 5) forudsættes at observationerne har en given fordeling (f.eks. normalfordelingen). For at sikre at forudsætningerne er opfyldt er det derfor hensigtsmæssigt at checke dette. I andre sammenhænge kan det være nyttigt at sammenligne et nyt datasæt med et ældre (eksempelvis eksamenskaraktererne ved et universitetskursus for to år).

Ved sammenligning af fordelingerne af to stokastiske variable hvoraf den ene (eller begge) er implicit givet ved et antal stikprøver, er det nødvendigt at gruppere stikprøverne i kvanter. Hvis kvanterne er meget små (mange kvanter) vil de fleste kvanter kun indeholde ganske få (eller ingen) observationer. Hvis kvanterne er meget store vil de fleste observationer ofte falde i meget få kvanter. I begge tilfælde vil en sammenligning af fordelingerne være vanskelig. Der findes ingen faste regler for opdeling af værdiområdet i kvanter. En tommelfingerregel for kvantstørrelsen er at der skal falde mindst 5 observationer i hvert kvant. Typisk samles observationerne fra de to haler af fordelingen i to "hale-kvanter". Antallet af observationer i disse bør være små (5-10). Vi skal tage denne diskussion op igen under omtalen af den såkaldte χ^2-test (udtales ki i anden). Først skal vi gennemgås Kolmogorov-Smirnov-testen.

Kolmogorov-Smirnov test

Kolmogorov-Smirnov testen er en særlig simpel form for sammenligning af en empirisk fordeling $F_x(x)$ og en teoretisk (kontinuert) fordeling $F_y(y)$. Bemærk at vi her benytter de akkumulerede fordelingsfunktioner (og ikke tæthedsfunktionerne). Lad n være antallet af observationer af x. $F_x(t)$ angiver da antallet af observationer med værdi mindre end eller lig t, divideret med n. Teststørrelsen er:

$$D = \max\{|F_x(x) - F_y(y)|\} \tag{C.7}$$

Under antagelsen at de to fordelinger er ens, kan signifikansniveauet for forkastning af hypotesen approksimeres ved funktionen:

$$Q(\lambda) = 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2\lambda^2} \tag{C.8}$$

Det kan vises at værdien $Q(\sqrt{n}D)$ er lig sandsynligheden for at den teoretiske værdi af D er større end den observerede teststørrelse. Hvis $Q(\sqrt{n}D)$ derfor er stor (tæt på 1), kan hypotesen om ens fordelinger accepteres.

χ^2 - goodness of fit test

χ^2-testen er en af de oftest benyttede test, dels fordi den er let at anvende, dels pga. metodens generalitet. Ved sammenligning af en empirisk fordeling, baseret på M stikprøver, og en teoretisk kontinuert fordeling foretages kvantificeringen af den teoretiske fordeling først. Antallet af teoretisk beregnede observationer skal summere op til det faktiske antal M. I denne beregning benyttes ofte en approksimation til sandsynlighedsmassen i hvert kvant. Lad K være antallet af kvanter, og lad n_i være antal teoretisk beregnede observationer i den i'te kvant. De faktiske observationer grupperes nu i de K kvanter. Lad N_i være antal faktiske observationer i den i'te kvant. Teststørrelsen for hypotesen, at de to fordelinger er ens er da:

$$\chi^2_{obs} = \sum_{i=1}^{K} \frac{(N_i - n_i)^2}{n_i} \tag{C.9}$$

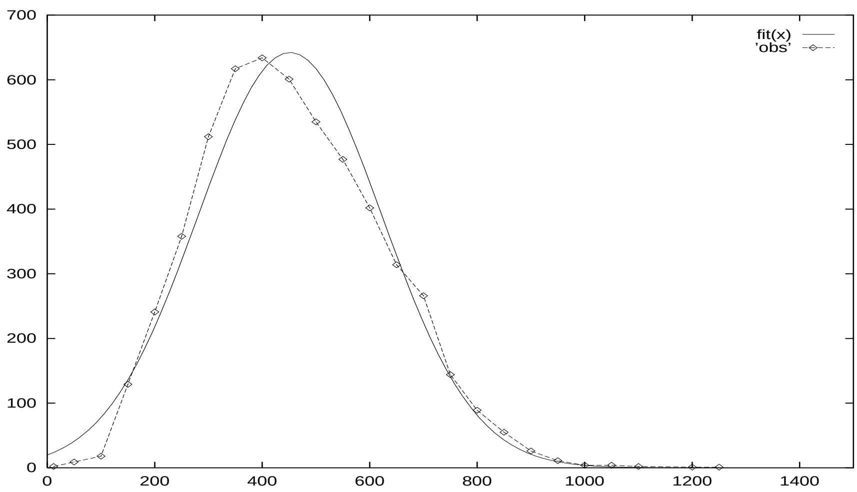
Det kan vises at teststørrelsen tilnærmelsesvis er χ^2-fordelt med $K-1$ frihedsgrader hvis $n_i > 5$. Dette beror på, at addenderne i summen tilnærmelsesvist vil være kvadratet på en normalfordelt størrelse med middelværdi 0 og spredning 1. Hvis de to fordelinger er meget forskellige ses at teststørrelsen vil blive stor. Vi vil derfor forkaste hypotesen om ens fordelinger (på et givet signifikansniveau) hvis størrelsen χ^2 er stor. Valg af lille signifikansniveau betyder her at kravet til lighed mellem de to fordelinger er stort.

Eksempel

På 5512 forskellige tidspunkter er antallet af processer, der afvikles på maskinen "Embla" registreret. Vi vil modellere antal processer som en stokastisk variabel x. Det teoretiske værdiområde for x er $[0, 1500]$, hvor den øvre grænse svarer til størrelsen af procestabellen. I praksis afvikles sjældent under 20 og sjældent over 1000 processer. En opremsning af de 5512 observationer af x vil være for voldsom her, i stedet er frekvenserne vist i nedenstående skema for en grovere kvantificering af værdiområdet.

processer	frekvens	processer	frekvens
0-25	2	676-725	266
26-75	9	726-775	144
76-125	78	776-825	89
126-175	129	826-875	55
176-225	241	876-925	26
226-275	358	926-975	11
276-325	512	976-1025	4
326-375	617	1026-1075	4
376-425	634	1067-1125	2
426-475	601	1125-1175	0
476-525	535	1176-1225	1
526-575	477	1226-1275	1
576-625	402	1276-1325	0
626-675	314	1326-1500	0

Middelværdien og spredningen af de 5512 observationer kan beregnes til $\mu = 451.8$ og $\sigma = 171.2$. Ved modellering af den empiriske fordeling ville det nok være hensigtsmæssigt at forsøge med en Poisson fordeling eller en negativ binomialfordeling. Imidlertid skal vi her teste om den empiriske fordeling kan antages at stemme overens med en normalfordeling med parametrene (451.8, 171.2). Et plot af observationerne og normalfordelingen er vist nedenstående.



Figur C.1:

Ved beregning af de teoretiske værdier viser det sig at disse er små for antal processer over 1000. Det vælges derfor at kvanter for værdier herover slås sammen. Iøvrigt lader den i tabellen viste opdeling rimelig. Nedenstående tabel viser de $K = 21$ endelige intervaller, størrelserne n_i, N_i, samt bidraget til χ.

processer	N_i	n_i	$\frac{(N_i-n_i)^2}{n_i}$
0-25	2	23.7	19.87
26-75	9	41.0	24.98
76-125	78	77.9	0.00
126-175	129	136.0	0.36
176-225	241	218.2	2.38
226-275	358	321.2	4.22
276-325	512	434.3	13.90
326-375	617	539.2	11.23
376-425	634	614.6	0.61
426-475	601	643.4	2.79
476-525	535	618.4	11.25
526-575	477	545.8	8.67
576-625	402	442.4	3.69
626-675	314	329.2	0.70
676-725	266	225.0	7.47
726-775	144	141.2	0.06
776-825	89	81.3	0.73
826-875	55	43.0	3.35
876-925	26	20.9	1.24
926-975	11	9.3	0.31
976-	12	6.0	6.00

Det ses at især de første to kvanter bidrager til χ^2 teststørrelsen, der bliver $\chi^2 = 123.81$. Denne teststørrelse er ret stor. Da der ialt er 21 kvanter bliver antallet af frihedsgrader lig 20. 0.99-fraktilen af $\chi(20)$ er lig 37.57. Da vores teststørrelse er større kan vi ikke engang på dette konfidensniveau acceptere hypotesen: at antallet af processer er normalfordelt. Det skal bemærkes at en anden kvantificering af fordelingerne ville give en anden (og måske mindre) teststørrelse. Antallet af observationer er imidlertid så stort, at de nok burde stemme bedre overens med den teoretiske normalfordeling hvis de virkelig var genereret af en sådan.

Eksempel slut

Ved sammenligning af to empiriske fordelinger er metoden stort set som beskrevet ovenfor. Blot er teststørrelsen nu:

$$\chi^2 = \sum_{i=1}^{K} \frac{(N_i - n_i)^2}{N_i + n_i} \tag{C.10}$$

Antallet af frihedsgrader er, som før. Det er væsenligt at kvantificeringen af de to sæt af observationer er ens, samt at disse er normeret således at $\sum n_i = \sum N_i$.

C.1.2 Ikke-parametrisk korrelation

Et væsenligt problem ved brug af den sædvanlige korrelationskoefficient er, at denne ikke er robust. Hvis alle punkter $(x[i], y[i])$, $i = 1, 2, \ldots, N$, bortset fra ét, ligger perfekt på en ret linie, kan værdien af $|r|$ være meget mindre end 1. Ikke parametrisk korrelation (også kaldet rank korrelation) er en robust metode, hvor hvert punkt $(x[i], y[i])$ erstattes af et nyt $(nx[i], ny[i])$, således at $nx[i]$ hhv. $ny[i]$ er indices for $x[i]$ hhv. $y[i]$ i de sorterede lister af koordinatværdier. De oprindelige koordinatværdier ignoreres herefter helt. *Spearmans rank-order korrelationskoefficient* er defineret ved:

$$r_S = \frac{\sum_i (nx[i] - \mu_{nx})(ny[i] - \mu_{ny})}{\sqrt{\sum_i (nx[i] - \mu_{nx})^2}\sqrt{\sum_i (ny[i] - \mu_{ny})^2}} \tag{C.11}$$

hvor μ_{nx} hhv. μ_{ny} er middelværdierne af indekstallene $nx[i]$ hhv. $ny[i]$. Lad nu:

$$t = r_S \sqrt{\frac{N - 2}{1 - r_S^2}} \tag{C.12}$$

Det kan vises at t er fordelt som Students t-fordeling, med $N - 2$ frihedsgrader.

Eksempel
Betragt nedenstående 12 punktpar:

i	1	2	3	4	5	6	7	8	9	10	11	12
x	101	102	202	203	305	405	407	508	709	810	910	911
nx	1	2	3	4	5	6	7	8	9	10	11	12
y	102	101	203	503	202	705	504	805	906	907	908	908
ny	2	1	4	5	3	7	6	8	9	10	11	12

Ved udregning fås $r_S = 0.965$ og $t \approx 12.21$. Ved opslag i Students t-fordeling for 10 frihedsgrader fås, for 0.99-fraktilen, grænseværdien 2.76. Vi kan altså med meget stor pålidelighed acceptere hypotesen, at der er en lineær korrelation mellem variablene x og y. Den sædvanlige korrelationskoefficient har, i dette tilfælde, også en høj værdi, nemlig $r = 0.90$.

Eksempel slut

C.2 Stokastiske processer

En stokastisk proces er en én eller flerdimensional følge af indicerede stokastiske variable. For éndimensionelle processer $x[i]$ betegner indices typisk en (ækvidistant samplet) tidsangivelse. For todimensionale processer $x[i][j]$, også kaldt *stokastiske felter*, tænkes ofte på billeder, hvor værdien (intensiteten) i hvert billedelement modelleres ved en stokastisk variabel. Modellering af indicerede størrelser vha. stokastiske processer er meget anvendte inden for signalanalyse og billedbehandling.

En stokastisk proces $x[i]$, $i = 1, 2, \ldots$ kaldes *stationær* (eng. wide sense stationarity) hvis middelværdien $m[i]$ og variansen $\sigma^2[i]$ af x er konstant (som funktion af i). Hvis x er flerdimensional erstattes kravet om ens varians, med et krav om at kovariansmatricen er konstant.

En stokastisk proces kaldes uafhængig hvis sandsynligheden for til tidspunktet (stedet) i at observere værdien $x = t$ er uafhængig af alle tidligere observerede værdier. Dette kan udtrykkes:

$$p(x[i] \ \mid \ x[i-1], x[i-2], \ldots, x[0]) \ = \ p(x[i]) \tag{C.13}$$

Processen $x[i]$ kaldes en *Markov process af n'te orden* hvis sandsynligheden for at observere værdien $x[i] = t$ til tiden i kun er afhængig af de n sidst observerede værdier:

$$p(x[i] \ \mid \ x[i-1], x[i-2], \ldots, x[0]) \ = \ p(x[i] \ \mid \ x[i-1], \ldots, x[i-n]) \tag{C.14}$$

Mange fænomener kan modelleres godt ved markovprocesser. I billeder er intensiteten i et givet billedelement ofte tæt på intensiteten i nabo billedelementerne, hvorimod korrelationen med intensiteter i billedelementer, der ligger langt fra hinanden, er forsvindende. Antagelsen om markovegenskaben benyttes derfor ofte i signal og billedanalyse, samt (som det beskrives i en anden del af kurset) ved kodning baseret på informationsteori.

Der findes en meget omfangsrig litteratur, såvel som et meget stort antal anvendelser, af stokastiske (markov) processer. Vi skal her kun omtale de såkaldte *Markov random fields* (MRF), og deres relation til *Gibbs random fields* (GRF) ved det såkaldte *Hammersley-Clifford teorem.*

Et markov random field (MRF) er et todimensionelt felt $x[i][j]$ af stokastiske variable, om hvilken markovbetingelsen er at:

$$\begin{aligned} & p(x[i][j] = t_{ij} \ \mid \ x[k][l] = t_{kl}, i \neq k, j \neq k) \\ = \ & p(x[i][j] = t_{ij} \ \mid \ x[k][l] = t_{kl}, i \neq k, j \neq k, (k,l) \in \mathcal{N}) \end{aligned} \tag{C.15}$$

hvor $\mathcal{N}$ er et naboområde, der skal opfylde visse egenskaber (som vi ikke skal omtale her). Et typisk anvendt naboområder er bestemt ved de fire naboer: $(i-1,j)$, $(i+1,j)$, $(i,j-1)$, og $(i,j+1)$. Et sådant MRF siges at være af orden 1. Hvis de 8 naboer benyttes siges feltet at være af orden 2.

Et Gibbs random field (GRF) mht. et naboområde $\mathcal{N}$ er defineret ved:

$$p(x[i][j]=t) \;=\; \frac{1}{Z}\; exp(-U(t)/T) \tag{C.16}$$

hvor Z er en normaliseringskonstant (så integralet af p bliver 1), hvor T er en parameter kaldet *temperaturen*, og hvor $U(t)$ er en *energifunktion*, der har formen:

$$U(t) \;=\; \sum_k V_k(t) \tag{C.17}$$

hvor $V_k(t)$ kaldes en *potentialfunktion*. Antallet af potentialfunktioner svarer typisk til antallet af bånd, der begrænser friheden af det stokastiske felt. Om $V_k(t)$ gælder at den kun afhænger af værdierne af x i naboområdet $\mathcal{N}$ (dvs. en markovegenskab). Vi skal senere i noterne se anvendelser af GRF's, der især benyttes i mange metoder til stokastisk optimering. Styrken ved formalismen er, at den tillader opskrivning af en sandsynlighedsfunktion, der har et særdeles attraktivt udseende, når blot de nødvendige energifunktioner kan formuleres. Det sidste er ofte forholdsvist let.

Det er ikke umiddelbart klart, at MRF's har noget som helst at gøre med GRF's. Der eksisterer imidlertid en sætning, kaldet *Hammersley-Clifford teoremet*, der siger, at et stokastisk felt er en GRF hvis og kun hvis det er et MRF. De to formuleringer er altså ækvivalente.

C.3 Estimation af parametre

I dette afsnit introduceres til to generelle principper for estimation af parametre $\boldsymbol{\theta}$ på basis af et sæt af observationer $\mathcal{O}$. Det er underforstået at observationerne ønskes modelleret (vides at kunne modelleres) ved en model, der er fuldstændigt specificeret ved parametrene $\boldsymbol{\theta}$. I *maksimum likelihood estimation* betragtes parametrene som størrelser, der er faste, men ukendte. Estimationen består i, på basis af observationerne, at finde de værdier, der maksimerer sandsynligheden for at se de fakstisk registrerede observationer. I en anden metode, *Bayes estimation* betragtes parametrene som stokastiske variable, der hver er specificeret ved en a priori fordeling. Ved brug af Bayes regel, samt viden om de faktisk registrerede observationer, omformes disse til en a posteriori sandsynlighed for parametrene. Ved maksimering af dette udtryk bestemmes de parameterværdier, der givet observationerne og den a priori viden har maksimal sandsynlighed. Vi skal senere i noterne se anvendelse af maksimum likelihood estimations metoden ved lineær regression. Bayes estimations metoden skal vi benytte intensivt når vi skal diskutere metoder til klassifikation.

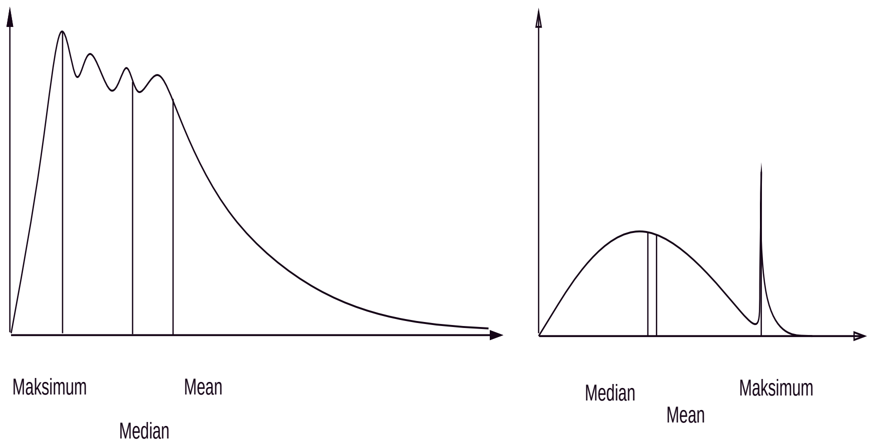
C.3.1 Maksimun likelihood estimation

I det følgende antages at mængden af observationer (stikprøver) kan skrives $\mathcal{O} = \{x_1, x_2, \ldots, x_n\}$, og at disse er udtrukket uafhængigt. Dette betyder at vi kan opskrive den betingede sandsynlighed for at iagttage de registrerede observationer givet sættet af de (endnu ukendte) parametre $\boldsymbol{\theta}$ som:

$$L(\boldsymbol{\theta}) \;=\; p(\mathcal{O}|\boldsymbol{\theta}) \;=\; \prod_{i=1}^{n} p(x_i|\boldsymbol{\theta}) \tag{C.18}$$

Størrelsen $L(\boldsymbol{\theta}) = p(\mathcal{O}|\boldsymbol{\theta})$, betragtet som funktion af $\boldsymbol{\theta}$, kaldes *likelihoodfunktionen* for $\boldsymbol{\theta}$. Den vektor $\hat{\boldsymbol{\theta}}$ der maksimerer likelihoodfunktionen kaldes et maksimum likelihood estimat.

Som indskud, skal der her bemærkes, at et valg af maksimum for likelihood funktionen på ingen måde er den eneste mulige strategi. To andre mulige estimater er middelværdien og medianværdien.



Figur C.2: Illustration af to fordelinger, med angivelse af maksimum, middelværdi, og median.

Der findes intet universelt bedste estimet. Det vil altid være op til problemløseren at definere hvad der skal forstås ved en optimal løsning. Hvis likelihood funktionen er pæn (konveks mv.) vil maksimumsværdien imidlertid ofte give det intuitivt korrekte estimat.

Det er ofte hensigtsmæssigt at arbejde med logaritmen til likelihood funktionen. Da logaritmefunktionen er monotont voksende vil maksimum af log-likelihood funktionen også være maksimum af likelihood funktionen selv. I mange praktiske anvendelser antages at likelihood funktionen er pæn, herunder konveks, og differentiabel. I dette tilfælde kan vi anvende metoder fra differentialregningen til at bestemme positionen af maksimum estimatet. Lad gradientoperatoren være defineret ved:

$$\nabla_{\boldsymbol{\theta}} \;=\; [\frac{\partial}{\partial\theta_1}, \frac{\partial}{\partial\theta_2}, \ldots, \frac{\partial}{\partial\theta_p}]^t \tag{C.19}$$

hvor p er antal parametre. Da er gradienten af log-likelihood funktionen givet ved:

$$\nabla_{\boldsymbol{\theta}} L \;=\; \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \, log \; p(x_i|\boldsymbol{\theta}) \tag{C.20}$$

og maksimum likelihood kriteriet er at $\nabla_{\boldsymbol{\theta}} L = 0$.

Eksempel
Antag at likelihood fordelingen er endimensional og normal. $\boldsymbol{\theta}$ bestemmer altså to parametre, middelværdien $\theta_1 = \mu$ og variansen $\theta_2 = \sigma^2$. Hvert element i log-likelihood funktionen ses at være:

$$log \; p(x_i|\boldsymbol{\theta}) \;=\; -\frac{1}{2} log(2\pi\theta_2) \;-\; \frac{(x_i - \theta_1)^2}{2\theta_2}$$

Gradienten af denne størrelse er:

$$\nabla_{\boldsymbol{\theta}} log \; p(x_i|\boldsymbol{\theta}) \;=\; \begin{bmatrix} \frac{x_i - \theta_1}{\theta_2} \\ \frac{1}{2\theta_2}\left(\frac{(x_i-\theta_1)^2}{\theta_2} \;-\; 1\right) \end{bmatrix}$$

Af disse udtryk fremkommer ligningerne:

$$0 \;=\; \frac{1}{\hat{\theta}_2} \sum_{i=1}^{n} (x_i \;-\; \hat{\theta}_1)$$

$$0 \;=\; \sum_{i=1}^{n} \frac{(x_i - \hat{\theta}_1)^2}{\hat{\theta}_2^2} \;-\; \sum_{i=1}^{n} \frac{1}{\hat{\theta}_2}$$

Af disse udtryk kan maksimum likelihood estimaterne $\hat{\theta}_1 = \hat{\mu}$ og $\hat{\theta}_2 = \hat{\sigma}^2$ let udledes:

$$\hat{\mu} \;=\; \frac{1}{n} \sum_{i=1}^{n} x_i$$

$$\hat{\sigma}^2 \;=\; \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2$$

Et fuldstændigt tilsvarende resultat kan udledes i det multivariate tilfælde, hvor parametrene er middelværdivektoren og kovariansmatricen. Som det fremgår af de tidligere afsnit i noterne er ovenstående estimat af variansen (og det tilsvarende for kovariansmetricen) et biased estimat. Selv om forskellen mellem estimaterne er lille for stor n, så illustrerer ovenstående at et maksimum likelihood estimat ikke nødvendigvis er unbiased.

Eksempel slut

C.3.2 Bayes estimation

Udgangspunktet i en Bayes estimation er kendskabet til de a priori sandsynligheder $p(\boldsymbol{\theta})$ for parametervektoren $\boldsymbol{\theta}$. Parametervektoren $\boldsymbol{\theta}$ betragtes altså som en stokastisk vektor. Essentielt i Bayes estimation er anvendelsen af Bayes regel. Denne siger at den a posteriori sandsynlighed for parametrene $\boldsymbol{\theta}$ givet de registrerede observationer $\mathcal{O}$ samt de a priori sandsynligheder $p(\boldsymbol{\theta})$ er givet ved:

$$p(\boldsymbol{\theta}|\mathcal{O}) \;=\; \frac{p(\mathcal{O}|\boldsymbol{\theta}) \;\; p(\boldsymbol{\theta})}{p(\mathcal{O})} \tag{C.21}$$

hvor $p(\mathcal{O})$ er sandsynligheden for netop at observere de registrerede observationer. Maksimum af dette udtryk, som funktion af $\boldsymbol{\theta}$, kaldes *maksimum aposteriori estimation* (MAP). Da nævneren i udtrykket er uafhængig af parametervektoren er denne blot en konstant, og kan udelades af optimeringen. Maksimum a posteriori estimationen er derfor bestemt ved:

$$\hat{\boldsymbol{\theta}}_{MAP} \;::\; \max_{\boldsymbol{\theta}} \{p(\mathcal{O}|\boldsymbol{\theta}) \;\; p(\boldsymbol{\theta})\} \tag{C.22}$$

Det er vigtig at huske, at $p(\boldsymbol{\theta})$ er ukendt som værdi betragtet, men antages at have en kendt form, f.eks. en normalfordeling med kendt middelværdi og varians. Udtryket $p(\mathcal{O}|\boldsymbol{\theta})$ er en funktion af $\boldsymbol{\theta}$ og er, som sådan fuldstændig kendt. Som kommenteret i sidste afsnit er maksimumsværdien af den aposteriori fordeling ikke det eneste selektionskriterium for valget af den optimale løsning. Som alternativ kan middelværdien eksempelvis benyttes.

Eksempel

Lad $\mathcal{O}$ bestå af n punkter $(x_1, x_2, \ldots, x_p, y)$ opnåede ved stikprøvetagning således at observationerne er uafhængige. Lad modellen af de observerede data være, at der findes en lineær relation mellem y og x'erne, således at der for alle observationer $i = 1, 2, \ldots, n$ gælder at:

$$y^i \;=\; a_0 \;+\; a_1 x_1^i \;+\; \cdots a_p x_p^i$$

hvor parametervektoren $\boldsymbol{\theta} = (a_0, a_1, \ldots, a_p)$. Lad for givet valg af $\boldsymbol{\theta}$, r_i være defineret ved tallet:

$$r_i(\boldsymbol{\theta}) \;=\; r_i \;=\; y^i \;-\; (a_0 + a_1 x_1^i + \cdots + a_p x_p^i)$$

og antag at $p(\mathcal{O}|\boldsymbol{\theta}) = p(r_i)$ er normalfordelt med middelværdi 0 og ukendt spredning σ.

Antag yderligere at vi ingen a priori viden har om fordelingen af parameterværdierne. I dette tilfælde er det måske rimeligt at antage at alle parameterværdier er lige sandsynlige, dvs. at $p(\boldsymbol{\theta})$ er en konstant, og derfor kan udelades af maksimeringen. I dette tilfælde vil MAP-estimatet stemme overens med maksimum likelihood estimatet. Da observationerne er uafhængige er MAP-estimatet givet ved den værdi $\hat{\boldsymbol{\theta}}$, der maksimerere:

$$h(\boldsymbol{\theta}) \;=\; \prod_{i=1}^{n} e^{-\frac{r_i^2(\boldsymbol{\theta})}{2\sigma^2}}$$

Da logaritmefunktionen er monontont voksende, og parameteren σ er konstant ses, at $\hat{\boldsymbol{\theta}}$ er bestemt ved minimum af udtrykket:

$$\hat{\boldsymbol{\theta}}_{MAP} \;::\; \min_{\boldsymbol{\theta}} \left\{ \sum_{i=1}^{n} r_i^2(\boldsymbol{\theta}) \right\}$$

Som beskrevet i Kapitel 3 i noterne kan $\hat{\boldsymbol{\theta}}$ findes af dette udtryk ved anvendelse sædvanlig lineær algebra (Mindste Kvadraters Metode).

Eksempel slut

I ovenstående eksempel antog vi, at de a priori sandsynligheder var konstanter. Dette var begrundet i at der ikke var noget forhåndskendskab til fordelingen af parameterværdierne til rådighed. Dette valg ses ofte foretaget i praksis (med lignende begrundelse). Der skal imidlertid indtrængende advares mod sådanne begrundelser. Hvis parametervektoren antages blot lidt anderledes fordelt, er det muligt at estimatet vil være radikalt forskelligt. Hvis det ikke af teoretisk vej er muligt at bestemme de a priori sandsynligheder kan den eneste sunde vej frem være at estimere fordelingerne eksperimentelt. Hvis parametervektoren har en kendt parametrisk form (med ukendte parameterværdier), er en anden mere avanceret mulighed, at bestemme fordelingen løbende ved indlæring. Vi skal ikke her diskutere sådanne metoder.

Eksempel

I stereoanalyse af billeder betragtes to billeder af samme scene. Hvert synligt punkt P_{3D} i scenen afbildes i to punkter, $P_L = (x_L, y_L)$ i det "venstre" billede og $P_R = (x_R, y_R)$ i det "højre" billede. De to billedpunkter siges at korrespondere. Givet de to billeder er der ikke umiddelbart klart hvorledes punkterne i disse skal parres således at alle punktpar korresponderer. Vi skal i dette eksempel se et eksempel på en metode til at estimere korrespondancerne. Dette svarer til at bestemme to (diskrete) funktioner d_x og d_y således at:

$$\begin{aligned} x_R &= x_L + d_x(x_L, y_L) \\ y_R &= y_L + d_y(x_L, y_L) \end{aligned}$$

Funktionerne d_x hhv. d_y kaldes den horisontale hhv. vertikale disparitet. Det er ofte muligt at optage billederne således at den verikale disparitet $d_y()$ er konstant lig 0. Dette skal vi antage i det følgende. Lad billederne være af dimension $N \times N$. For hvert af de N^2 punkter i det venstre billede er der således N mulige kandidater til løsning af korrespondanceproblemet, hvis vi antager at hvert venstre-punkt har netop ét korresponderende højre-punkt (denne antagelse er noget problematisk). Uden yderligere oplysninger er der derfor N^{2N} forskellige løsninger til problemet. I praksis findes to ofte benyttede a priori antagelser, der begrænser løsningsrummet.

- Intensiteten I_L og I_R i to korresponderende punkter P_L og P_R er tilnærmelsesvis ens. Bemærk at denne antagelse ikke holder for blanke eller spejlende materialer. På grund af støj i billedoptagelsen kan forskellen i intensitet ofte modelleres ved en normalfordeling med middelværdi 0. Støjen i to punkter antages at være uafhængig.

- Dispariteten d_x er tilnærmelsesvis ens i nabopunkter. Denne antagelse er rimelig hvis scenefladen er glat, men holder ikke de steder hvor afstanden til sceneobjekterne ændrer sig diskontinuert.

Nedenstående gives en simpel skitse til hvorledes et estimat af dispariteten d_x kan findes. Vi skal modellere $d_x(x_L, y_L)$ som en stokastisk variabel. Sandsynligheden for at denne antager værdien D i punktet (x_L, y_L) givet intensiteterne I_L i (x_L, y_L) of I_R i $(x_L + D, y_L)$ kan opskrives som:

$$p(D|I_L, I_R) \;=\; \frac{p(I_L, I_R, D)}{p(I_L, I_R)} \;=\; \frac{p(I_R|D, I_L)\; p(D, I_L)}{p(I_R, I_L)} \;=\; \frac{p(I_R|D, I_L)\; p(D|I_L)}{p(I_R|I_L)}$$

Der er ikke i almindelighed grund til at tro at dispariteten er afhængig af intensiteten I_L. Derfor er $p(D|I_L) = p(D)$. Nævneren i udtrykket er uafhængig af D. Den værdi, der maksimerer $p(I_R|D, I_L)\; p(D)$ vil derfor også maksimere venstresiden af ovenstående udtryk. Vi skal nu udnytte den første af de tidligere nævnte antagelser, at forskellen i intensitet i korresponderende punkter er normalfordelt med middelvædi 0. Dette betyder at:

$$p(I_R|D, I_L) \;=\; p([I_R(x_L + D, y_L) - I_L(x_L, y_L)] = t) \;=\; \frac{1}{\sqrt{2\pi}\sigma}\; exp\left(-\frac{t^2}{2\sigma^2}\right)$$

Da støjen i to punkter er antaget uafhængig er den samlede betingede sandsynlighed for at se det højre billede giver det venstre billede samt disparitetsfunktionen lig produktet (over alle billedpunkter) af ovenstående udtryk.

For at kunne udnytte den anden antagelse, at funktionen d_x er glat, er det hensigtsmæssigt at modellere denne som et markov random field. Dette betyder, jvf. Hammersley-Clifford teoremet, at vi kan skrive:

$$p(D = t) \;=\; \frac{1}{Z}\; exp[-U(t)/T]$$

hvor Z er normaliseringskonstanten $Z = \sum_t exp[-U(t)/T]$, hvor T er en konstant (temperaturen), og hvor energifunktionen U har formen:

$$U(t) \;=\; \sum_C V_C(t)$$

hvor V_C kun afhænger af elementerne i en naboomegn. Vi skal her benytte den simplest tænkelige naboomegn $\mathcal{N}$, nemlig de fire nærmeste naboer. Herved fås et MRF af orden 1. Da d_x antages omtrentlig ens i nabopunkterne er det oplagt at definere:

$$V_C(t) \;=\; \sum_{i \in \mathcal{N}} [t - t_i]^2$$

Størrelsen V_C måler således den lokale afvigelse fra antagelsen om konstant disparitet. I dette tilfælde har vi:

$$p(D) \;=\; \frac{1}{Z}\; exp\left[-\frac{1}{T}\sum_{k=-1,1}\;\sum_{l=-1,1}(d_x(i,j) - d_x(i+k, j+l))^2\right]$$

I det følgende kan vi ignorere konstanten Z i optimeringen. Lad endvidere $\lambda = \frac{T}{2\sigma^2}$. Da er aposteriori fordelingen $p(D|I_L, I_R)$ proportional med:

$$\prod_{(i,j)\in[N\times N]} exp\left(-\frac{\lambda}{T}[I_R(i + d_x(i,j), j) - I_L(i,j)]^2\right) exp\left(-\frac{1}{T}\sum_{k=-1,1}\;\sum_{l=-1,1}[d_x(i,j) - d_x(i+k, j+l)]^2\right)$$

Idet logaritmefunktionen er monotont voksende, og temperaturen T er en konstant ses, at maksimum aposteriori estimatet af $d_x(i,j)$ er bestemt ved minimum af udtrykket:

$$H(d_x) \;=\; \sum_{(i,j)\in[N\times N]} \Big(\lambda[I_R(i+d_x(i,j),j) - I_L(i,j)]^2 \;+\; Glat(i,j)\Big)$$

hvor

$$\begin{aligned} Glat(i,j) \;&=\; \quad [d_x(i,j) - d_x(i-1,j-1)]^2 \;+\; [d_x(i,j) - d_x(i-1,j+1)]^2 \\ &=\; + \, [d_x(i,j) - d_x(i+1,j-1)]^2 \;+\; [d_x(i,j) - d_x(i+1,j+1)]^2 \end{aligned}$$

Hvad vi har vist er, at vi givet nogle relativt vage forudsætninger har været i stand til at opstille et relativt kompakt udtryk for kvaliteten H af en disparitetsfunktion d_x. Udtrykket angiver, for hver punkt i det venstre billede, den straf en given disparitetsværdi, vil bidrage med. Udtryket er ikke-negativt og lig 0 netop når forudsætningerne er eksakt opfyldt. Den funktion d_x, der minimerer funtionalen H er et Bayes estimat af en mest sandsynlige disparitetsfunktion.

Det ses, at udtrykket H er parametriseret af konstanten λ. Denne parameter angiver et mål for troværdigheden af antagelsen om ens intensitet i korresponderende punkter i forhold til antagelsen om glathed af disparitetsfunktionen. En mulighed er at fastlægge en bestemt værdi af λ fra starten (fundet ved andre metoder eller gættet). En anden strategi er itterativt at approksimere en løsning således at λ i starten er lille, og vokser med iterationsnummeret. Til start kræves således stor glathed af d_x. Senere i processen slækkes på dette krav i forhold til kravet om overensstemmelse af intensiteterne i korresponderende punkter. Det kan bemærkes at funktionen H (for hver fast værdi af λ) meget vel kan have mange minima. Dette afhænger helt af billederne I_L og I_R. Vi skal ikke her diskutere hvorledes minimeringen kan foretages, men vil udskyde dette til et senere afsnit i noterne.

Eksempel slut

C.4 Opgaver

1. I et større eksperiment blev visse sammenhænge mellem køn (d/p), matematisk evne (M) og sproglig evne (S) undersøgt. I eksperimentet deltog 277 piger og 318 drenge. Ved analyse af de eksperimentelle data fandt man at:

   - 40 % opnåede en M-værdi større end eller lig 8.
   - Blandt de drenge, der havde en S-værdi mindre end 5, opnåede halvdelen en M-værdi mindre end 8.
   - 60 % af de der havde en M-værdi mindre end 8 opnåede en S-værdi mindre end 5.
   - 70 % af drengene opnåede en S-værdi mindre end 5.

   Givet at $M < 8$, og at $S < 5$ bedes du angive sandsynligheden for at en forelagt observation stammer fra en dreng.

Appendix D

Robuste estimatorer

Som beskrevet i noterne findes der en række robuste estimatorer af middelværdi, af varians etc. Nedenstående beskrives nogle robuste metoder til løsning af overbestemte lineære ligningssyatemer.

D.1 Iterativt vægtede mindste kvadraters metode

I visse situationer vides på forhånd at nogle ligninger $y_i = \mathbf{x}_{i.}\mathbf{a}$ er vigtigere end andre. I disse tilfælde kan ligningerne vægtes ved at gange såvel højre- som venstresiden med en konstant w_i. Lad $\mathbf{W}$ være givet ved $diag(w_1, w_2, \ldots, w_k)$. Det vægtede lineære ligningssystem er nu givet ved:

$$\mathbf{r} \;=\; \mathbf{W}\left(\mathbf{y} \;-\; \mathbf{X}\mathbf{a}\right) \tag{D.1}$$

Dette system kan løses præcis som tidligere beskrevet ved at benyttte $\mathbf{X}_W = \mathbf{W}\mathbf{X}$ og $\mathbf{y}_W = \mathbf{W}\mathbf{y}$. Problemet ved vægtede metoder er at vælge vægtene. Hvis det ikke på forhånd er givet, at visse målinger er mere troværdige (eller på anden måde vigtigere) end de øvrige, kunne man forestille sig at bestemme vægtene således at observationer, der ellers ville give anledning til store residualer, vægtes lidt, mens de observationer, hvor r_i er lille vægtes mere. Dette svarer til at erstatte funktionen ρ i ligning (3.15) med en funktion, der vokser langsommere end r_i^2. En sådan estimator kaldes en *M-estimator* (navnet kommer fra: “Maksimum likelihood type estimate”). Det kan vises, at sammenhængen mellem vægtene w_i og funktionen ρ er givet ved:

$$w_i \;=\; \frac{1}{r_i}\frac{\partial \rho(r_i)}{\partial r_i} \tag{D.2}$$

Vælges $\rho(r) = r^2/2$, svarende til 2-normen, fås $w = 1$. Vælges $\rho(r) = |r|$, svarende til 1-normen, fås $w = 1/r$. Mere realistisk kunne man sætte en øvre grænse T på hvor meget en stor residual må påvirke løsningen:

$$\rho(r) \;=\; \max\left\{T, \frac{1}{2}r^2\right\} \;\;\Rightarrow\;\; w \;=\; \begin{cases} 1 & \text{hvis } r \leq \sqrt{2T} \\ 0 & \text{ellers} \end{cases} \tag{D.3}$$

En typisk vægtfunktion er kontinuert, symmetrisk og består (for positive r) af tre områder. Tættest ved 0 er mindste-kvadraters-området hvor vægten er 1. Længst væk fra 0 er outlier-området, hvor vægten er 0. Imellem er vægtfunktionen aftagende. En klassisk vægtfunktion er den såkaldte *Bi-weight* funktion:

$$w = \begin{cases} [1 - (\frac{r}{c\hat{\sigma}_r})^2]^2 & \text{hvis } |r| \leq c\hat{\sigma}_r \\ 0 & \text{ellers} \end{cases} \tag{D.4}$$

hvor $\hat{\sigma}_r$ er et robust estimat af spredingen af residualværdierne og c er en konstant. En anden klassisk estimator, kaldet *Lorentz-estimatoren* fremkommer ved valg af $\rho(r) = \frac{1}{2}\log(1 + r^2)$.

Ved implementation af vægtede mindste kvadraters metode ses, at et initielt estimat af residualerne er nødvendige for at kunne bestemme vægtene. Derfor implementeres disse metoder ofte ved en iterativ procedure. Først anvendes den uvægtede mindste kvadraters metode. Herved bestemmes $\mathbf{r}$ og $\hat{\sigma}$. Disse størrelser bruges til beregning af vægtene $\mathbf{w}$, og et nyt estimat af $\mathbf{r}$ kan nu udledes, og processen itereres. Stopkriteriet er ofte en kombination af et maksimalt øvre antal af iterationer og et mål for konvergens af metoden. Et typisk sådant mål er C^t:

$$\begin{aligned} C^t &= |E^t - E^{t-1}| \\ E^t &= \sqrt{\frac{\sum_{i=1}^{k} w_i^t(r_i^t)(r_i^t)^2}{\sum_{i=1}^{k} w_i^t(r_i^t)}} \end{aligned} \tag{D.5}$$

hvor superscript t angiver iterationsindeks. Hvis C^t falder under en vis størrelse, f.eks 10^{-3} stoppes processen.

M-estimatiorer er betydeligt mere rubuste (hvad angår afvigende y-værdier) end de rå mindste kvadraters metode, men har stadig et lavt nedbrudspunkt. Kvaliteten af det første fit er afgørende for metodens succes, og det er ikke let at afgøre, hvilken betydning forskellige valg af funktionen ρ har på resultatet. I de senere år har opmærksomheden derfor koncentreret sig om estimationsmetoder, der har højt nedbrudspunkt, og som ikke er følsomme over for startpunkt og parametervalg. En sådan metode er *mindste medianers metode*.

D.2 Mindste medianers metode

Mindste medianers metode fremkommer ved bestemmelse af de regressionsparametre der minimerer medianværdien af kvadratresidualerne.

$$\min_{\mathbf{a}} \left\{ med(r_i^2) \right\} \tag{D.6}$$

Da medianværdien er ufølsom over for numerisk store residualer har mindste medianers metode samme høje nedbrudspunkt 0.5 som medianværdien selv. Problemet ved metoden er primært, at denne er beregningsmæssigt meget tung. Nedenfor er skitseres en algoritme, der benytter udtømmende søgning. Lad k være antallet af observationer og n antallet af regressionsparametre. Lad M være den hidtil mindste værdi af medianen af residualerne, initielt en stor værdi.

For samtlige $\binom{k}{n}$ kombinationer af n data $(\mathbf{x}, y)$ blandt de k observationer:

1. Beregn sættet af regressionsparametre ved løsning af n ligninger med n ubekendte.
2. Beregn sættet af k residualværdier r_i, og medianen m af $|r_i|$.
3. Hvis $m < M$ registreres sættet af regressionsparametre, som det hidtil bedste, og M sættes lig m.

Som det ses er beregningskompleksiteten af ovenstående algoritme ganske overvældende. I appendix B er en væsentlig hurtigere metode, baseret på stokastisk sampling, illustreret. En anden observation er, at metoden ikke er særlig nøjagtig, idet det bedste sæt af regressionsparametre altid er bestemt således, at der er eksakt overensstemmmelse med n observationer. Mindste medianers metode benyttes derfor ofte udelukkende til at identificere outliers. Efterfølgende benyttes en binært vægtet mindste kvadraters metode (der vides at være optimal) på inliers. Vægtene bestemmes ved:

$$w_i = \begin{cases} 1 & \text{hvis } \frac{|r_i|}{\hat{\sigma}_r} \leq 2.5 \\ 0 & \text{ellers} \qquad \text{- outliers sorteres fra} \end{cases} \tag{D.7}$$

hvor r_i er mindste medianers residualerne og $\hat{\sigma}_r$ er bestemt ved ligning (1.15). Som det ses er metoden ikke følsom over for noget initielt estimat. Den eneste parameter i algoritmen er konstanten 2.5, og denne er ikke kritisk.

D.3 MDL-estimation

MDL er en forkortelse af *Minimum Description Length* [10], og betegner en metode, hvor antallet af bit, der er nødvendige for kodning af samtlige data ønskes minimeret. I forhold til andre metoder indgår kompleksiteten af den anvendte model i beregningen af beskrivelseslængden (DL). Den grundlæggende antagelse er, at den korteste beskrivelseslængde antages ved den mest hensigtsmæssige modellering. En meget stor fordel ved MDL-konceptet er, at det tillader at sammenligne kvaliteten af flere væsenligt forskellige modeller. Eksemplevis kan en model af en en-dimensional funktion baseret på en polynomieopløsning umiddelbart sammenlignes med én baseret på en frekvensopløsning. Simplere kan polynomie-modeller af forskellig orden sammenlignes mht. kodningseffektivitet. Dette er ikke let ved andre metoder. Den omtalte fordel opnås, ved at måle kompleksiteten af modellen og effektiviteten af denne (dvs. residualerne ved anvendelse af modellen på de forelagte data) i samme enheder, nemlig i bit. Jævnfør antagelsen (3.1) skal vi her nøges med at betragte lineære modeller med et fast antal n af parametre.

MDL-konceptet hører til en gren af teoridannelser inden for informationsteorien. I denne sammenhæng er problemstillingen at sende en datamængde over en kanal, f.eks. en telefonlinie. Det er her af oplagt interesse at minimere antallet af transmiterede bit. MDL-konceptet beskriver hvorledes dette antal udregnes. Væsentlig er at al information kodes. I dette afsnit skal vi kun diskutere MDL-konceptet perifert. Konkret ønsker vi at opstille et udtryk for antallet af bit, der skal til at kode informationen i et overbestemt lineært ligningssystem med n ubekendte. Tallet n antages at være kendt af "modtageren". Ved rå kodning af samtlige k ligninger skal der kodes $k(n+1)$ tal. Hvis fittet er perfekt (ingen outliers, og ingen støj) skal der kun kodes n tal. Denne repræsentation må derfor anses for bedre, fordi antallet af bit, der går til at kode n tal, er mindre end det, der går til at kode $k(n+1)$ tal.

For fast model (lineær model i n ubekendte) afhænger beskrivelseslængden DL (antallet af bit) af opdelingen af observationer i inliers og outliers, samt af den præcision hvormed modellen fitter til inliers.

$$DL_{total} \;=\; DL_{model} \;+\; DL_{residual} \;+\; DL_{outlier} \tag{D.8}$$

hvor DL_{model} er antallet af bit, der går til kodning af de n modelparametre, $DL_{residual}$ er antallet af bit, der går til kodning af residualværdierne, og $DL_{outlier}$ er antallet af bit, der går til kodning af eventuelle outliers. Hvis alle observationer er inliers, og hvis størrelsen DL_{model} er en konstant, dvs. er uafhængig af de faktisk estimerede parameterværdier, da vil den den minimale beskrivelseslængde være bestemt ved minimum for $DL_{residual}$. Hvis residualværdierne er normalfordelte med middelvædi 0, kan det vises, at MDL-værdien (dvs. $\min\{DL\}$) antages for det parametersæt der minimerer variansen af residualerne. I dette tilfælde stemmer løsningen derfor overens med løsningen ved mindste kvadraters metoder, og er derfor et maksimum likelihood estimat.

Algoritmen til beregning af den minimale beskrivelseslængde ligner meget algoritmen benyttet ved mindste medianers metode. Lad MDL initielt antage en stor værdi:

For samtlige $\binom{k}{n}$ kombinationer af n data $(\mathbf{x}, y)$ blandt de k observationer:

1. Beregn sættet af regressionsparametre ved løsning af n ligninger med n ubekendte.
2. Beregn sættet af k residualværdier r_i.
3. For $i = 0$ to $i = k - n$
   - Opdel data i en gruppe af de i outliers, der har de største numeriske residualer, og en gruppe af $k - i$ inliers.
   - Beregn beskrivelseslængden DL^i_{total}.
   - Hvis $DL^i_{total} < MDL$ registreres sættet af regressionsparametre, som det hidtil bedste, og MDL sættes lig DL^i_{total}.

Det ses, at beregningskompleksiteten af ovenstående algoritme er værre end for mindste medianers metode, fordi beskrivelseslængden skal beregnes $(k - n + 1)$ gange hver gang, der beregnes en ny residualvektor. Som for mindste medianers metode kan kompleksiteten nedbringes hvis stokastisk sampling benyttes.

For at kunne anvende metoden i regression behøver vi ikke at foretage nogen egentlig kodning. Derimod skal vi, for at kunne benytte ovenstående algoritme, kunne beregne (et godt estimat af) kodelængden DL_{total}. Inliers kan kodes ved modelparametrene $\mathbf{a}$, samt residualerne r_i ved modelleringen. For outliers, der ikke kan modelleres godt må de oprindelige data $\{(\mathbf{x_i}, y_i)\}$, kodes.

Ved beregning af beskrivelseslængderne er det vigtigt at benytte en optimal kodning. I modsat fald vil kodemetoden bestemme hvilke parametre, der udvælges. Der findes mange metoder til at indkode en sekvens af tal. Vi skal her kun omtale præfikskodningen (der er nærmere beskrevet i en anden del af kurset). Ligeledes skal vi ikke her diskutere hvornår denne indkodning er optimal. Hvis en stokastisk heltallig variabel v er uniformt fordelt inden for et interval $[0 : L]$, kan det vises, at den gennemsnitlige kodelængde af v er givet ved $-\log_2(L^{-1})$. Ved kodning af et reelt tal v ved en endelig binær streng er det nødvendigt at specificere den nøjagtighed, hvormed værdien skal angives. Lad $\epsilon = 2^{-q}$ angive denne nøjagtighed. Antallet af bit, der går til kodning af v er omkring $\log_2(|v|/\epsilon)$, hvis fortegnsbitten ignoreres. Ved kodning af flere tal ved en præfikskodning adderes bit til beskrivelse af længden af kodningen (således at det i en lang binær streng kan afgøres hvornår et tal slutter og et andet starter). Den samlede beskrivelseslængde $DL(v)$ for v kan vises [10] at være begrænset af:

$$1 \ + \ \log_2(\frac{|v|}{\epsilon}) \ \leq \ DL(v) \ \leq \ 1 \ + \ \log_2(\frac{|v|}{\epsilon}) \ + \ 2 \ \log_2 \log_2(\frac{|v|}{\epsilon}) \tag{D.9}$$

hvor det første led er kodningen af fortegnbitten. Hvis hver værdi kodes med forskellig præcision er det ydeligere nødvendigt at addere bit $DL(q_i)$ til kodning af konstanten q_i. Hvis $q_i = q$ er konstant (fælles for alle størrelser v) behøver vi kun addere $DL(q)$ én gang.

Ved kodning af inliers skal parametrene a_j, samt residualværdierne kodes. Det vil føre for vidt her at diskutere hvorledes ϵ bestemmes for parameterværdierne. Ved kodning af residualværdierne for inliers er det ofte rimeligt at antage at disse er normalfordelte med middelværdi 0 og spredning σ. Den information (eng. self-information), som en stokastisk variabel r med fordeling $p(r)$ indeholder, er fra informationsteorien kendt som $-\log_2 p(r)$. Dette er altså det mindste antal bit, hvormed vi kan kode r uden at tabe information. Med udgangspunkt i ligning (3.11) fås heraf at beskrivelseslængden for p inliers er:

$$\begin{aligned} DL_{residual} &= -\log_2\left(\prod_{inliers} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{r_i}{\sigma})^2}\right) \\ &= p \ \log_2(\sqrt{2\pi}\sigma) \ + \ \frac{1}{2 \ ln(2)} \sum_{inliers} (\frac{r_i}{\sigma})^2 \end{aligned} \tag{D.10}$$

Da $r_i = y_i - \mathbf{Xa}$ er præcisionen ϵ af r_i lig præcisionen af y_i. Da σ er i samme skala (størrelsesorden) som r_i er det rimeligt at antage at præcisionen af σ også er ϵ. Vi får derfor at:

$$\begin{aligned} DL_{residual} &= p \ \log_2(\sqrt{2\pi}\frac{\sigma}{\epsilon}) \ + \ \frac{1}{2 \ ln(2)} \sum_{inliers} (\frac{r_i/\epsilon}{\sigma/\epsilon})^2 \\ &= p \ \left[\log_2(\frac{\sigma}{\epsilon}) \ + \ \frac{1}{2}\log_2(2\pi e)\right] \end{aligned} \tag{D.11}$$

hvor vi har udnyttet reglerne for omskrivning mellem logaritmer med forskelligt grundtal, at den stokastiske variabel $\frac{r_i}{\sigma}$ har middelværdi 0 og varians 1, og at antallet af inliers er stort.

Ved beregning af kodelængden for outliers skal værdien af hver modelvariable $\mathbf{x_i}$ og hver observerbar variabel y_i kodes. For outliers er ϵ bestemt ved nøjagtigheden af de foreliggende data. Desuden skal det kodes hvilke ligninger i, der er klassificeret som outliers.

Ved anvendelse af MDL-konceptet i estimation, ønsker vi at finde det sæt af parametre $\mathbf{a}$, der minimerer DL_{total}. De konstanter i dette udtryk, der er uafhængig af opdelingen af observationerne i inliers og outliers, samt af $\mathbf{a}$ of dermed $\mathbf{r}$, har ingen indflydelse på minimumspunktet, og kan ignoreres. Væsentlige spørgsmål, som vi efterlader, er valg af indkodningsmetode, og bestemmelse af præcisionen af modelparametrene.

Indeks