Ekstrakt af gamle noter i MønsterGenkendelse

S. I. Olsen

Dette ekstrakt er tænkt til hjælp for studerende på kurset *Introduktion til billedbehandling* 2005. Ekstraktet er foretaget med henblik på at understøtte kursusnoterne inden for grundlæggende statistike begreber og mindste kvadraters metode. Desuden er appendix med nogle grundlæggende begreber inden for lineær algebra og parameterestimation bevaret.Studerende bedes udvise overbærenhed dels med noternes form (de er ret gamle), dels med den manglende sammenhæng (grundet ekstrakten). Ekstraktet er ment som en hjælp, men indeholder afsnit, der (i anden sammenhæng) på ingen måde erstatte et egenligt undervisningsmateriale. Som for andre noter forudsættes det at læseren har et modent (operationelt) forhold til matematik.

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.
$\mathbf{x}_{.j}$	Søjlevektor af matrice.
$\mathbf{x}_{i.}$	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.

Indhold

Kapitel 1

Statistisk Dataanalyse

I dette kapitel introduceres til nogle grundlæggende statisktisk begreber og estimationsmetoder. 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.

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 forbrændt benzin ved kørsel på Jagtvej). For $n = 1$ er $F(x)$ en reel funktion af en varia-

bel. 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.

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 $g : \mathcal{R} \to \mathcal{R}$ 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:

1 $\sigma^2(x) \geq 0$
2 $\sigma^2(x) = E(x^2) - [E(x)]^2$
3 $\sigma^2(ax) = a^2\sigma^2(x)$ for $a \in \mathcal{R}$
4 $\sigma^2(x + a) = \sigma^2(x)$ for $a \in \mathcal{R}$

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 øges 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 medianfilteret 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	5	3

mål	σ	$MeanAD$	$MedianAD$	$Skew$	$SKEW_{kvartil}$	$Kurt$
værdi	3.51	3.08	1.48	1.92	1	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 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:

$$r \;=\; \frac{\sum_i (x[i] - \hat{\mu}_x)(y[i] - \hat{\mu}_y)}{(n-1)\;\hat{\sigma}_x\;\hat{\sigma}_y} \tag{1.17}$$

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 (se evt. [6]).

1.6 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:

$$Cov(x, y) \;=\; E([x - E(x)][y - E(y)]) \tag{1.18}$$

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

$$\begin{aligned}
\mathbf{C} &= E([\mathbf{x}-\mathbf{m_x}][\mathbf{x}-\mathbf{m_x}]^t) \qquad (1.19)\\
&= E\left(\begin{bmatrix} x_1-m_1 \\ x_2-m_2 \\ \vdots \\ x_n-m_n \end{bmatrix} [(x_1-m_x),(x_2-m_2),\ldots,(x_n-m_x)]\right)\\
&= \begin{bmatrix} E([x_1-m_1][x_1-m_1]) & E([x_1-m_1][x_2-m_2]) & \cdots & E([x_1-m_1][x_n-m_n]) \\ E([x_2-m_2][x_1-m_1]) & E([x_2-m_2][x_2-m_2]) & \cdots & E([x_2-m_2][x_n-m_n]) \\ \vdots & & \ddots & \vdots \\ E([x_n-m_n][x_1-m_1]) & E([x_n-m_n][x_2-m_2]) & \cdots & E([x_n-m_n][x_n-m_n]) \end{bmatrix}\\
&= \begin{bmatrix} VAR(x_1) & COV(x_1,x_2) & \cdots & COV(x_1,x_n) \\ COV(x_2,x_1) & VAR(x_2) & \cdots & COV(x_2,x_n) \\ \vdots & & \ddots & \vdots \\ COV(x_n,x_1) & COV(x_n,x_2) & \cdots & VAR(x_n) \end{bmatrix} \qquad (1.20)
\end{aligned}$$

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:

$$\mathbf{C} \;=\; E(\mathbf{x}\mathbf{x}^t) \;-\; E(\mathbf{x})\mathbf{m_x}^t \;-\; \mathbf{m_x}E(\mathbf{x}^t) \;+\; \mathbf{m_x}\mathbf{m_x}^t \;=\; \mathbf{S} \;-\; \mathbf{m_x}\mathbf{m_x}^t \qquad (1.21)$$

hvor

$$\mathbf{S} \;=\; E(\mathbf{x}\mathbf{x}^t) \;=\; \begin{bmatrix} E(\mathbf{x}_1\mathbf{x}_1) & \cdots & E(\mathbf{x}_1\mathbf{x}_n) \\ \vdots & & \vdots \\ E(\mathbf{x}_1\mathbf{x}_n) & \cdots & E(\mathbf{x}_n\mathbf{x}_n) \end{bmatrix} \qquad (1.22)$$

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

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

$$\mathbf{\Lambda} \;=\; \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & & \sigma_n \end{bmatrix} \qquad (1.23)$$

og

$$\mathbf{R} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{1n} & 1 & & \vdots \\ \vdots & & \ddots & \vdots \\ \rho_{1n} & \cdots & & 1 \end{bmatrix} \tag{1.24}$$

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

$$c_{ij} = \begin{cases} \sigma_i^2 & \text{hvis } i = j \\ \rho_{ij}\sigma_i\sigma_j & \text{hvis } i \neq j \end{cases} \tag{1.25}$$

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 (1.17). 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 (1.21) 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:

$$\hat{\mathbf{C}} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x} - \hat{\mathbf{m}})(\mathbf{x} - \hat{\mathbf{m}})^t \tag{1.26}$$

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:

$$\hat{\mathbf{C}} = \begin{bmatrix} 3.200 & 3.667 & 1.666 \\ 3.667 & 6.783 & 2.075 \\ 1.666 & 2.075 & 6.329 \end{bmatrix} \tag{1.27}$$

$$= \begin{bmatrix} 1.789 & 0 & 0 \\ 0 & 2.604 & 0 \\ 0 & 0 & 2.516 \end{bmatrix} \begin{bmatrix} 1.000 & 0.787 & 0.370 \\ 0.787 & 1.000 & 0.317 \\ 0.370 & 0.317 & 1.000 \end{bmatrix} \begin{bmatrix} 1.789 & 0 & 0 \\ 0 & 2.604 & 0 \\ 0 & 0 & 2.516 \end{bmatrix}$$

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:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}|\mathbf{C}|^{1/2}} \; exp[-\frac{1}{2}(\mathbf{x}-\mathbf{m})\mathbf{C}^{-1}(\mathbf{x}-\mathbf{m})^t] \tag{1.28}$$

$$= \frac{1}{(2\pi)^{n/2}|\mathbf{C}|^{1/2}} \; exp[-\frac{1}{2}d^2(\mathbf{x})] \tag{1.29}$$

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 (1.28) til:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\sigma_1\sigma_2\cdots\sigma_n} \; exp\left[-\frac{1}{2}\sum_{i=1}^{n}\frac{(\mathbf{x}-\mathbf{m})^2}{\sigma_i^2}\right] \tag{1.30}$$

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.

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:

$$\begin{aligned} y_1 &= \phi_1(x_1, x_2, \ldots, x_n) \\ y_2 &= \phi_2(x_1, x_2, \ldots, x_n) \\ &\vdots \qquad\qquad\qquad \vdots \\ y_n &= \phi_n(x_1, x_2, \ldots, x_n) \end{aligned} \tag{1.31}$$

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:

$$f_y(\mathbf{y}) \;=\; f_x(\psi(\mathbf{y}))\left|\frac{\partial(\mathbf{x})}{\partial(\mathbf{y})}\right| \tag{1.32}$$

for alle $\mathbf{y}$ i værdimængden for ϕ. Bemærk at der tages numerisk værdi af den anden faktor i (1.32). Størrelsen

$$\frac{\partial(\mathbf{x})}{\partial(\mathbf{y})} \;=\; det\left\{\frac{\partial x_i}{\partial y_j}\right\}_{i,j=1,..,n} = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \dots & \frac{\partial x_1}{\partial y_n} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} & \dots & \frac{\partial x_2}{\partial y_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \dots & \frac{\partial x_n}{\partial y_n} \end{vmatrix} \tag{1.33}$$

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

Kapitel 2

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 $j = 0 \ldots n-1$. I resten af dette kapitel skal vi antage, at afhængigheden er lineær, dvs. at afhængigheden kan skrives:

$$y \;=\; a_0 \;+\; \sum_{j=1}^{n-1} a_j x_j \;= a_0 \;+\; a_1 x_1 \;+\; a_2 x_2 \;+\; \ldots \;+\; a_{n-1} x_{n-1} \;=\; \mathbf{x}_j \mathbf{a} \qquad (2.1)$$

hvor $\mathbf{x}_j = (1, x_1, x_2, \ldots, x_{n-1})$, og hvor $\mathbf{a} = (a_0, a_1, \ldots, a_{n-1})^t$. 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:

$$\begin{aligned} y_1 &= a_0 + a_1 x_{11} + a_2 x_{12} + \ldots + a_{n-1} x_{1(n-1)} \\ y_2 &= a_0 + a_1 x_{21} + a_2 x_{22} + \ldots + a_{n-1} x_{2(n-1)} \\ \vdots &\quad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ y_k &= a_0 + a_1 x_{k1} + a_2 x_{k2} + \ldots + a_{n-1} x_{k(n-1)} \end{aligned} \tag{2.2}$$

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:

$$\mathbf{y} = \mathbf{Xa} \tag{2.3}$$

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:

$$\mathbf{r} = \mathbf{y} - \mathbf{Xa} \tag{2.4}$$

Afvigelsen $r_i = \mathbf{x}_{i\cdot}\mathbf{a}$ 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.

2.1 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:

$$\begin{aligned} r^2 &= \mathbf{r}^t\mathbf{r} = ||\mathbf{y} - \mathbf{Xa}||^2 \\ &= (\mathbf{y} - \mathbf{Xa})^t(\mathbf{y} - \mathbf{Xa}) \\ &= \mathbf{y}^t\mathbf{y} - \mathbf{y}^t\mathbf{Xa} - \mathbf{a}^t\mathbf{X}^t\mathbf{y} + \mathbf{a}^t\mathbf{X}^t\mathbf{Xa} \\ &= \mathbf{y}^t\mathbf{y} - 2\mathbf{y}^t\mathbf{Xa} + \mathbf{a}^t\mathbf{Ca} \end{aligned} \tag{2.5}$$

hvor $\mathbf{C} = \mathbf{X}^t\mathbf{X}$, dvs. at elementerne $c_{ij} = \mathbf{x_{\cdot i}^t x_{\cdot j}}$ 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:

$$\mathbf{b} \;=\; \mathbf{Ca}^* \;=\; \mathbf{X}^t\mathbf{Xa}^* \tag{2.6}$$

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 (2.5). Lad $r^2(\mathbf{a}^*)$ angive værdien af kvadratresidualet i ligning (2.5) når den entydige løsning $\mathbf{a}^*$ benyttes. Lad $\mathbf{h}$ være en vilkårlig vektor i $\mathcal{R}^n$. Da gælder:

$$\begin{aligned} r^2(\mathbf{a}^* + \mathbf{h}) \;&=\; \mathbf{y}^t\mathbf{y} \;-\; 2\mathbf{y}^t\mathbf{X}(\mathbf{a}^* + \mathbf{h}) \;+\; (\mathbf{a}^* + \mathbf{h})^t\mathbf{C}(\mathbf{a}^* + \mathbf{h}) \\ &=\; \mathbf{y}^t\mathbf{y} \;-\; \mathbf{b}^t\mathbf{a}^* \;+\; \mathbf{h}^t\mathbf{Ch} \end{aligned} \tag{2.7}$$

I ligning (2.7) 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 (2.4) overens med løsningen til normalligningen (2.6). Hvis $\mathbf{C}$ er positiv definit er den invertibel. Mindste kvadraters løsning til det oprindelige ligningssystem kan derfor udtrykkes ved:

$$\hat{\mathbf{a}} \;=\; (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} \tag{2.8}$$

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 (2.3) 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:

$$y_i \;=\; a_0 \;+\; a_1x_i \;+\; a_2x_i^2 \;+\; a_3x_i^3 \tag{2.9}$$

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

$$\mathbf{X} \;=\; \begin{pmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^2 & x_2^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_k & x_k^2 & x_k^3 \end{pmatrix} \tag{2.10}$$

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 (2.8).

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

2.1.1 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:

$$p(r_i) \;=\; \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{y_i - \mathbf{x}_{i.}\mathbf{a}}{\sigma})^2} \tag{2.11}$$

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

$$p(\mathbf{r}) \;=\; \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^k \prod_{i=1}^{k} e^{-\frac{1}{2}(\frac{y_i - \mathbf{x}_{i.}\mathbf{a}}{\sigma})^2} \tag{2.12}$$

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:

$$f(\mathbf{a}) \;=\; p(\mathbf{r}) \tag{2.13}$$

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

$$k\ \log(\sqrt{2\pi}\sigma) \;+\; \sum_{i=1}^{k} \frac{1}{2}(\frac{y_i - \mathbf{x}_{i.}\mathbf{a}}{\sigma})^2 \tag{2.14}$$

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:

$$\min \sum_{i=1}^{k} \rho(\mathbf{r}) \;=\; \min \sum_{i=1}^{k} r_i^2 \tag{2.15}$$

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.

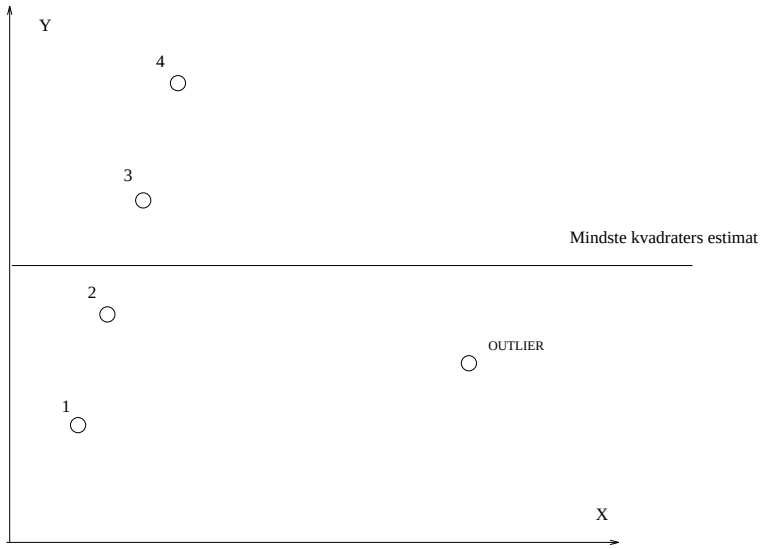
årsagen til, at mindste kvadraters metode, på trods af ovennævnte forhold, er så populær, er dels, at der findes meget effektive beregningsmetoder til denne metode, dels *De store tal lov* fra sandsynlighedsregningen. Denne siger, at hvis fejlen på en given observation kan beskrives som summen af uendelig 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.

2.2 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 og en ligefordeling. I dette tilfælde er mindste kvadraters metode ikke længere optimal.

Betragt tilfældet illustreret i Figur 2.2, 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 2.1: 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 iørigt 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 2.2 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:

$$\mathbf{X} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 10 \end{pmatrix} \quad \mathbf{X}^t\mathbf{X} = \begin{pmatrix} \sum 1 & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix} = \begin{pmatrix} 5 & 20 \\ 20 & 130 \end{pmatrix} \tag{2.16}$$

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

$$(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t = \frac{1}{250} \begin{pmatrix} 110 & 90 & 70 & 50 & -70 \\ -15 & -10 & -5 & 0 & 30 \end{pmatrix} \tag{2.17}$$

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:

$$\mathcal{Z} \;=\; \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_k, y_k)\} \tag{2.18}$$

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:

$$B(m, \mathcal{Z}) \;=\; \sup_{\mathcal{Z}_m} ||T(\mathcal{Z}_m) - T(\mathcal{Z})|| \tag{2.19}$$

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:

$$N_T \;=\; \min\left\{\frac{m}{k} \mid B(m, \mathcal{Z}) = \infty\right\} \tag{2.20}$$

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:

$$\mathbf{r} \;=\; \mathbf{y} - \hat{\mathbf{y}} \;=\; \mathbf{y} - \mathbf{X}\hat{\mathbf{a}} \;=\; \mathbf{y} - \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} \;=\; (\mathbf{I} - \mathbf{H})\mathbf{y} \tag{2.21}$$

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 $\hat{\mathbf{y}} = \mathbf{Hy}$ 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.

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 (2.11). Typisk identificeres observationer som potentielle outliers hvis $|r_i| > 2.5\ \sigma_r$.

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.

2.3 RANSAC

En standard metode ved estimation baseret på data med outliers er RANSAC *RANdom SAmple Consensus*. Om tiden tillader vil jeg skrive noter om dette. Indtil da, brug google.

Litteratur

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

[2] C.M. Bishop: *Neural Networks for Pattern Recognition*; Clarendon Press 1995.

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

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

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

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

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

Bilag 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:

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

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 $||\cdot|| : \mathcal{V} \rightarrow \mathcal{R}^+$, der opfylder at:

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

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 $\mathcal{V} \times \mathcal{V} \rightarrow \mathcal{R}$ sådan at:

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

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:

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

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

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

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.

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

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 $\mathbf{Xy} = \mathbf{0}$ 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 $\mathbf{Xy}_1 = \mathbf{Xy}_2$. 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:

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

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:

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

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 $det\ (\mathbf{XY}) = det\ \mathbf{X}\ det\ \mathbf{Y}$, 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:

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

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:

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

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:

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

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 $\mathbf{A}$ og $\mathbf{B}$ kaldes *ækvivalente* hvis der findes en ikke-singulær matrice $\mathbf{R}$ således at:

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

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

A.4 Lineære ligningssystemer

Lad $\mathbf{X}$ være en reel $k \times n$ matrice, og $\mathbf{y}$ en vektor med k elementer. Da $\mathbf{X}$ 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 $\mathbf{y}$. 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 $\mathbf{X}$ er lineært uafhængige. Hvis systemet er underbestemt findes uendelig mange løsninger, svarende til et underrum af $\mathcal{R}^n$. Hvis søjlerne i $\mathbf{X}$ er lineært uafhængige er dimensionen af underrummet lig $n - k$. Hvis systemet er overbestemt findes der ingen løsninger $\mathbf{a}$ til dette. Givet et $\mathbf{a}$ kan fejlen, eller resudualen $\mathbf{r}$ 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 $\mathbf{r}$ (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. [6] for egentlige algoritmebeskrivelser.

A.5.1 Gauss-elimination

Antag et kvadratisk lineært ligningssystem $\mathbf{C}\mathbf{a} = \mathbf{b}$ 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:

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

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 $\mathbf{y} = \mathbf{Ax}$ 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:

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

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 $\mathbf{y} = \mathbf{Ax}$ være et overbestemt $k \times n$ ligningssystem, hvor rangen af $\mathbf{A}$ er lig n. *Normalligningen* fremkommer ved multiplikation med $\mathbf{A}^t$, dvs:

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

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 $\mathbf{A} = \mathbf{UWV}^t$:

$$\begin{bmatrix} \\ \mathbf{A} \\ \\ \end{bmatrix} = \begin{bmatrix} \\ \mathbf{U} \\ \\ \end{bmatrix} \cdot \begin{bmatrix} w_1 & & & \\ & w_2 & & \\ & & \ddots & \\ & & & w_n \end{bmatrix} \cdot \begin{bmatrix} \\ \mathbf{V}^t \\ \\ \end{bmatrix}$$

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:

$$\begin{aligned} \mathbf{x} &= \mathbf{V} \cdot [diag(r(w_i))] \cdot (\mathbf{U}^t \mathbf{y}) \\ r(w_i) &= \begin{cases} \frac{1}{w_i} & \text{hvis } w_i > T \\ 0 & \text{hvis } w_i \leq T \end{cases} \end{aligned}$$

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

Bilag B

Sandsynligheder og estimation

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

B.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 $p : \mathcal{S} \to \mathcal{R}$ 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. $p(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} p(A_i) \quad$ hvis der gælder at $\quad \forall i \neq j : A_i \cap A_j = \emptyset$

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:

$$p(A|B) \;=\; \frac{p(A \;\cap\; B)}{p(B)} \tag{B.1}$$

De to hændelser siges at være *uafhængige* hvis der gælder at:

$$p(A|B) \;=\; p(A) \tag{B.2}$$

Generelt gælder at:

$$p(A|B) \;=\; \frac{p(A \;\cap\; B)}{p(B)} \;=\; \frac{p(B \;\cap\; A)}{p(A)}\frac{p(A)}{p(B)} \;=\; \frac{p(B|A)\,p(A)}{p(B)} \tag{B.3}$$

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 $x : \Omega \to \mathcal{R}$, forstås den funktion p_x, der til en delmængde Π af $\mathcal{R}$ bestemmer sandsynligheden for at $x \in \Pi$, dvs.:

$$p_x(\Pi) \;=\; p(\{\omega | x(\omega) \in \Pi\}) \;=\; p(x^{-1}(\Pi)) \tag{B.4}$$

Fordelingsfunktionen F for en stokastisk variabel $x : \Omega \to \mathcal{R}$ er en funktion $F : \mathcal{R} \to [0,1]$ defineret ved:

$$F(t) \;=\; p(x \leq t) \;=\; p_x(]-\infty, t]) \tag{B.5}$$

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. $\lim_{x\to-\infty} F(x) = 0$
2. $\lim_{x\to\infty} F(x) = 1$
3. F er voksende
4. F er kontinuert fra højre
5. $\lim_{x\to t+} F(x) \;-\; \lim_{x\to t-} F(x) \;=\; p(x = t)$

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:

$$f(x) = F'(x) \quad \forall x \in \mathcal{R} \tag{B.6}$$

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 $F(x) = \int_{-\infty}^{x} f(t)dt$.

B.2 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.

B.2.1 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:

$$L(\boldsymbol{\theta}) \;=\; p(\mathcal{O}|\boldsymbol{\theta}) \;=\; \prod_{i=1}^{n} p(x_i|\boldsymbol{\theta}) \tag{B.7}$$

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.
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:

$$\nabla_{\boldsymbol{\theta}} \;=\; [\frac{\partial}{\partial\theta_1}, \frac{\partial}{\partial\theta_2}, \ldots, \frac{\partial}{\partial\theta_p}]^t \tag{B.8}$$

hvor p er antal parametre. Da er gradienten af log-likelihood funktionen givet ved:

$$\nabla_{\boldsymbol{\theta}} L \;=\; \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}}\; log\; p(x_i|\boldsymbol{\theta}) \tag{B.9}$$

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:

$$\begin{aligned} 0 \;&=\; \frac{1}{\hat{\theta}_2} \sum_{i=1}^{n} (x_i \;-\; \hat{\theta}_1) \\ 0 \;&=\; \sum_{i=1}^{n} \frac{(x_i - \hat{\theta}_1)^2}{\hat{\theta}_2^2} \;-\; \sum_{i=1}^{n} \frac{1}{\hat{\theta}_2} \end{aligned}$$

Af disse udtryk kan maksimum likelihood estimaterne $\hat{\theta}_1 = \hat{\mu}$ og $\hat{\theta}_2 = \hat{\sigma}^2$ let udledes:

$$\begin{aligned} \hat{\mu} \;&=\; \frac{1}{n} \sum_{i=1}^{n} x_i \\ \hat{\sigma}^2 \;&=\; \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})^2 \end{aligned}$$

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

B.2.2 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:

$$p(\boldsymbol{\theta}|\mathcal{O}) \;=\; \frac{p(\mathcal{O}|\boldsymbol{\theta})\;\; p(\boldsymbol{\theta})}{p(\mathcal{O})} \tag{B.10}$$

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:

$$\hat{\boldsymbol{\theta}}_{MAP} \;::\; \max_{\boldsymbol{\theta}} \{p(\mathcal{O}|\boldsymbol{\theta})\;\; p(\boldsymbol{\theta})\} \tag{B.11}$$

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.

Indeks