26 - INTERPOLASJON OG REGRESJON

Dersom du har skjønt projektiv lineæralgebra, blir det enklere å huske og forstå hvordan lineærregresjon funker. I sannsynlighetsregningen pakkes dette inn i mye komplisert notasjon, og statistikerne må gjøre det slik for å kunne si noe om usikkerheten knyttet til parameterestimeringen. Vi skal i denne økten fokusere på de geometriske ideene som ligger i bånn for matematikken.

La oss begynne med noe som kalles **interpolasjon**. Hvis du har $n+1$ punkter (z_i, x_i) i $\mathbb{R}^2$, der $z_j \neq z_k$ for alle $j \neq k$, vil det alltid være mulig å finne et reelt polynom

$$p(z) = a_n z^n + a_{n-1} z^{n-1} + ... + a_1 z + a_0$$

hvis graf går gjennom alle disse punktene, altså at

$$p(z_k) = x_k$$

for alle $0 \leq k \leq n$. Disse likningene utgjør et $(n+1) \times (n+1)$-likningssystem for koeffisientene a_k med totalmatrise

$$\left(\begin{array}{ccccc|c} z_0^n & z_0^{n-1} & ... & z_0 & 1 & x_0 \\ z_1^n & z_1^{n-1} & ... & z_1 & 1 & x_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ z_n^n & z_n^{n-1} & ... & z_n & 1 & x_n \end{array}\right)$$

Dette systemet kalles **vandermondesystemet**, og vi slipper faktisk å plages med det, for vi kan heller gjøre slik: For hvert punkt z_k, definerer vi et polynom

$$l_k(z) = \prod_{\substack{j=0 \\ j \neq k}}^{n} \frac{(z - z_j)}{(z_k - z_j)}.$$

1 Sjekk at polynomet $l_k(z)$ har orden n, og at det tilfredsstiller

$$l_k(z_m) = \begin{cases} 1 & \text{for } m = k \\ 0 & \text{for } m \neq k \end{cases}$$

Det er lett nå å se at

$$p_n(z) = \sum_{k=0}^{n} x_k l_k(z)$$

tilfredsstiller $p_n(z_k) = x_k$ for alle k. Polynomene l_k kalles **lagrangepolynomer**, og alt dette kalles **Lagranges interpolasjonsmetode**. Det er også ikke veldig vanskelig å se at det alltid finnes et entydig interpolasjonspolynom av maksimal grad n dersom $z_k \neq z_j$ for alle $k \neq j$. [1]

2 Finn et tredjegradspolyom som går gjennom punktene $(0,1)$, $(1,0)$, $(2,1)$ og $(3,2)$.

Polynominterpolasjon ligger til grunn for veldig mye forskjellig. En klassiker er numerisk integrasjon, og dette skal vi se på i neste uke. Nå skal vi se på **regresjon**. Dette betyr at man man prøver å interpolere et datasett med et polynom som har for lav orden, og da må man forholde seg til det overbestemte $(m+1) \times (n+1)$-systemet

$$\left(\begin{array}{ccccc|c} z_0^n & z_0^{n-1} & \dots & z_0 & 1 & x_0 \\ z_1^n & z_1^{n-1} & \dots & z_1 & 1 & x_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ z_m^n & z_m^{n-1} & \dots & z_m & 1 & x_m \end{array}\right)$$

der $m > n$.

4 Prøv å finne et annengradspolyom som går gjennom punktene $(0,1)$, $(1,0)$, $(2,1)$ og $(3,2)$.

[1] Hvis vi antar at det finnes to forskjellige polynomer p_n og q_n av grad n som interpolerer den samme tabellen, og evaluerer differansen $p_n - q_n$ i punktene x_k, ser vi at

$$p_n(x_k) - q_n(x_k) = 0 \qquad 0 \leq k \leq n.$$

Polynomet $p - q$ har maksimal grad n, og kan maksimalt ha n nullpunkter, så derfor må $p = q$.

Likningssystemet i oppgaven over ble

$$\begin{aligned} a_0 &= 1 \\ a_2 + a_1 + a_0 &= 0 \\ 4a_2 + 2a_1 + a_0 &= 1 \\ 9a_2 + 3a_1 + a_0 &= 2 \end{aligned} \qquad \text{eller} \qquad \left(\begin{array}{ccc|c} 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 4 & 2 & 1 & 1 \\ 9 & 3 & 1 & 2 \end{array}\right)$$

Hvordan fikse dette? Nå trenger vi noe som kalles minste kvadraters metode. Dette er en heuristisk teknikk for å finne en tilnærmet løsning til et lineært system med flere likninger enn ukjente. Dette kan gjøres på mange måter, men minste kvadraters metode er forholdsvis enkel å forstå, den er basert på en pen geometrisk ide, og metoden gir en generell følelse av velvære.

La oss si at A er en $m \times n$-matrise, at $\mathbf{v}$ er en ukjent kolonnevektor i $\mathbb{R}^n$ og $\mathbf{x}$ er gitt kolonnevektor i $\mathbb{R}^m$, og at vi ønsker å finne $\mathbf{v}$ slik at

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

når $m > n$. Dette går antagelig ikke, for det er jo flere likninger enn ukjente, og systemet vil ikke ha noen løsning med mindre $\mathbf{x}$ tilfeldigvis ligger i kolonnerommet til A. La oss heller finne den $\mathbf{v}$ som minimerer avstanden fra $A\mathbf{v}$ til $\mathbf{x}$. Dette kan du gjøre på to forskjellige måter. Jeg har allerede fortalt deg om den ene i oppgave 20 - 14.

5 Gjør oppgave 4 med teknikken du fikk i økt 20. Du må bruke Gram-Schmidts metode og få tak i en ortogonal basis for kolonnerommet til matrisen

$$\begin{pmatrix} 0 & 0 & 1 \\ 1 & 1 & 1 \\ 4 & 2 & 1 \\ 9 & 3 & 1 \end{pmatrix}$$

og så bruke projeksjon.
(Denne oppgaven er ganske hårete, så ikke ta den seriøst. Bruk heller normallikningene.)

Det finnes imidlertid en annen vei til Rom, som ikke går via en ortogonal basis for kolonnerommet til A. Alle er enige i at om kolonnerommet til A er whiteboardet i figuren nederst på siden og at vi ønsker at den røde vektoren skal være ortogonalt på dette kolonnerommet og at dette kan vi få til ved å kreve at $\mathbf{x} - A\mathbf{v}$ må stå ortogonalt på alle kolonnene i A. Da må vi prikke $\mathbf{x} - A\mathbf{v}$ med alle kolonnene i A, og dette får vi enkelt og greit til på matriseform ved å kreve at

$$A^T(A\mathbf{v} - \mathbf{x}) = \mathbf{0}.$$

Det er vanlig å skrive

$$A^T A\mathbf{v} = A^T\mathbf{x}$$

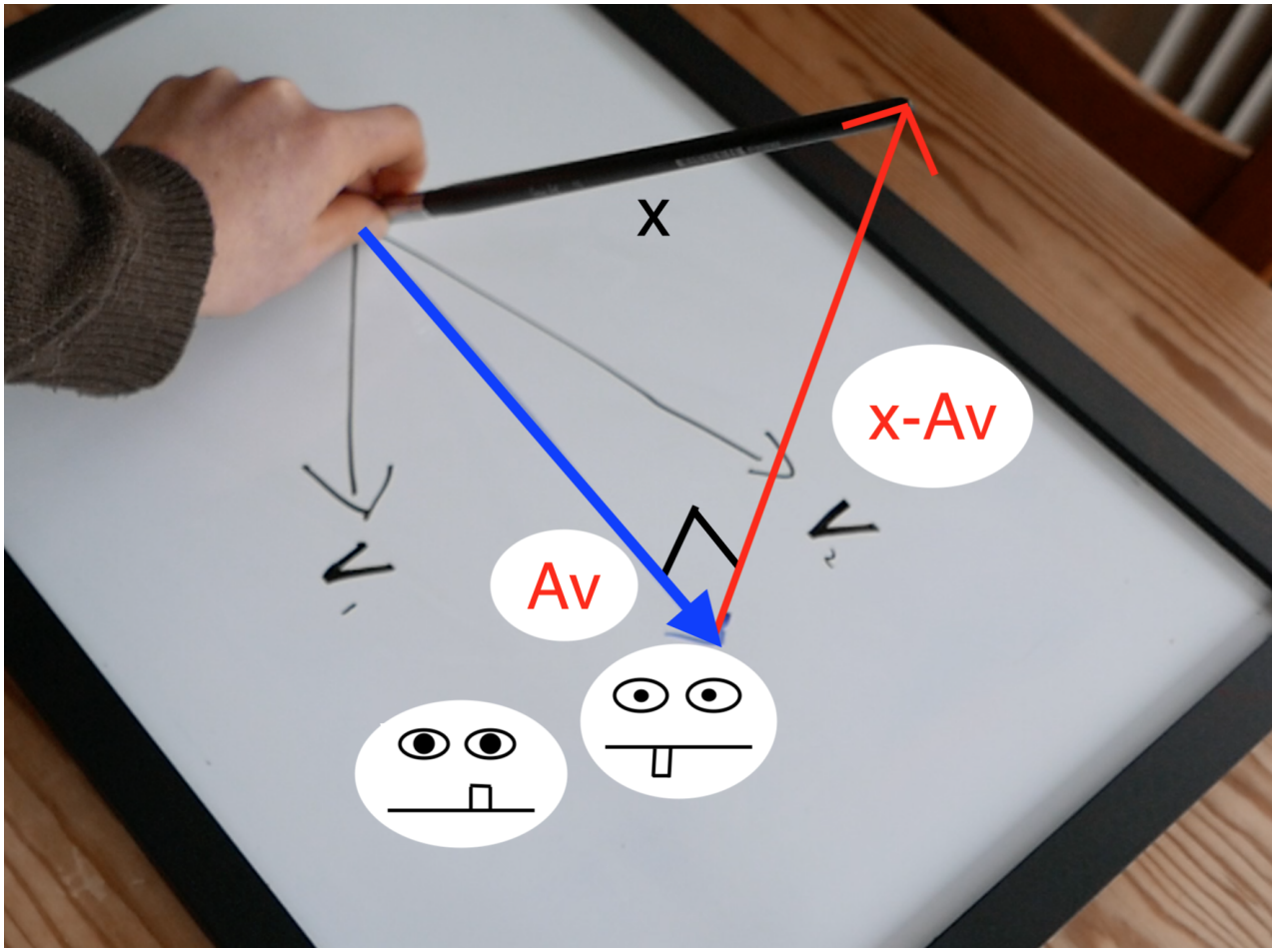
og dette er et $n \times n$-system som kalles **normallikningene**, Systemet har entydig løsning om kolonnene til A er lineært uavhengige, og løsningen $\mathbf{v}$ minimerer avstanden fra $A\mathbf{x}$ til $\mathbf{y}$ i $\mathbb{R}^n$.

6 Gjør oppgave 4 på denne måten.

Når statistikere snakker om **regresjonslinje**, mener de alt dette og at man bruker et førsteordens polynom $a_1 z + a_0$ og at man bruker minste kvadraters metode for å finne koeffisientene a_1 og a_0. Økonomer kaller det trendlinje. Husk at det er viktig å regne på usikkerhet, og det er derfor statistikere går litt grundigere til verks når de setter opp alt.

7 Finn regresjonslinjen til punktene $(0, 1)$, $(1, 0)$, $(2, 1)$ og $(3, 2)$.

8 Skriv en rutine som tar inn et datasett og produserer trendlinjen.



Dersom vi interpolerer med komplekse eksponensialfunksjoner istedet for polynomer, kalles det **diskret fourieromvending** (DFT). [2] Fourieromvending kommer i flere varianter, og vi har sett to varianter til nå: dersom du har et T-periodisk signal x, kan du stort sett skrive

$$x(t) = \sum_{n=-\infty}^{\infty} c_n e^{2\pi i n t/T} \qquad \text{der} \qquad c_n = \frac{1}{T}\int_{-T/2}^{T/2} x(t) e^{-2\pi i n t/T}\, dt.$$

og har du et ikkeperiodisk signal, kan du stort sett skrive

$$x(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} X(\omega) e^{i\omega t} d\omega \qquad \text{der} \qquad X(\omega) = \int_{-\infty}^{\infty} x(t) e^{-i\omega t}\, dt.$$

Diskret fourieromvending er den tredje varianten vi skal se på.

Nå er det litt forvirrende at normaliseringsfaktoren $1/T$ står på fourieromvendingen i det periodiske tilfellet, mens den tilsvarende faktoren $1/2\pi$ står på inversomvendingen i det ikkeperiodiske tilfellet, men slik er nå en gang konvensjonene. Når det gjelder diskret fourieromvending er notasjonen enda mere forvirrende, for profesjonelle diggsiggere bruker ikke diskret fourieromvendingsformler som likner på disse i det hele tatt. De gjør det som følger. La x være et diskret signal med N verdier $x(n)$. Da er det slik at dersom

$$X_k = \sum_{n=0}^{N-1} x(n) e^{-2\pi i n k/N} \qquad \text{er} \qquad x(n) = \frac{1}{N}\sum_{k=0}^{N-1} X_k e^{2\pi i n k/N}.$$

Dette blir bra greier med lite griseri. La oss gjenta oppgave 1.

2 Vis at at funksjonene $\phi_k : [0, N-1] \to \mathbb{C}$ gitt ved

$$\phi_k(n) = e^{2\pi i n k/N} \qquad \text{der} \qquad 0 \le k \le N-1$$

er ortogonale med hensyn på indreproduktet

$$(x, y) = \sum_{n=0}^{N-1} x(n)\overline{y(n)}$$

og utled at dersom

$$X_k = \sum_{n=0}^{N-1} x(n) e^{-2\pi i n k/N} \qquad \text{er} \qquad x(n) = \frac{1}{N}\sum_{k=0}^{N-1} X_k e^{2\pi i n k/N}.$$

for alle $n \in [0, N-1]$.

[2] "Discreet Fourier Transform" på engelsk.

Beregning av fourierkoeffisientene er et matrise-vektorprodukt, og dersom primtallsfaktoriseringen til N består av så mange små faktorer som mulig ($N = 2^m$ er det aller beste) går det veldig fort å beregne dem:
`https://en.wikipedia.org/wiki/Fast_Fourier_transform`
Dette kalles FFT (Fast Fourier Transform) og regnes som en av århundrets viktigste algoritmer, og mobiltelefonen din hadde aldri fungert uten denne. At fourierkoeffisientene kan beregnes raskt var kjent for Carl Friedrich Gauss, for han brukte visst diskret fourieromvending til å interpolere et datasett på jakt etter banene til asteroidene Pallas og Juno. Algoritmen har blitt gjenoppdaget mange ganger opp gjennom, men det var Cooley og Tukey som til slutt ble kreditert i 1965:
`https://en.wikipedia.org/wiki/Cooley-Tukey_FFT_algorithm`

10 Skriv opp matrise-vektorproduktet for $N = 7$ og $N = 8$. Se nøye på vandermondematrisene, og se om du ser noen forskjell. Hvis du ser veldig nøye etter, vil du kanskje skjønne hvordan Cooley og Tukey klarte å redusere antall beregninger fra størrelsesorden N^2 til $N \log N$.

Vi må selvfølgelig ha pytagoras, som sier at

$$\sum_{n=0}^{N-1} |x(n)|^2 = \frac{1}{N} \sum_{n=0}^{N-1} |X_k|^2$$

11 Jeg kunne bedt deg utlede dette, men du har jo allerede gjort det. Derfor slipper du det nå. Dette er fordelen med abstrakt matematikk.

Dette kalles ikke egentlig pytagoras, men derimot Plancherels teorem, jeg har bare kalt det pytagoras til nå siden det er det det er en generalisering av. En generalisering av Plancherel er Parsevals teorem, som sier at

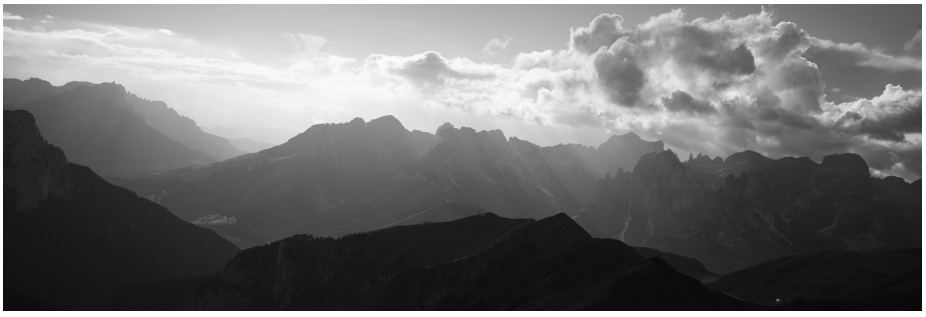
$$(x, y) = \frac{1}{N}(X, Y).$$

12 Klarer du denne?

Diskret fourieromvending brukes selvfølgelig til filtrering av digitale signal, og alle signaler er digitale nå til dags. Derfor må vi også ha konvolusjonsteoremet. Diggisiggfolk skriver for øvrig $x[n]$ og ikke $x(n)$. Diskret konvolusjon er

$$(x * y)[n] = \sum_{m=0}^{N-1} x[m] y\left[(n-m) \mod N\right].$$

13 Vis at den diskrete fouriertransformen til $x * y$ evaluert i k er $(X, Y)_k$.
Hint: $\mod N$ betyr "rest etter divisjon med N":
`https://en.wikipedia.org/wiki/Modulo`



UKENS NØTTER

Det går fint å regregere med andre typer funksjoner enn polynomer. Teknikken er det samme, sett opp likningssystem og bruk minste kvadraters metode. Jeg var på en hytte i Kragerø en gang og badet i en sirkulær badestamp med et sirkulært høl i bånn. Når jeg skulle tømme den, målte jeg vannstanden etterhvert som stampen tømtes, og fikk tabellen i margen. En enkel modell for tømmingen er

t (s)	h (cm)
0	48
10	38
20	28
30	20
40	14
50	8

$$\dot{h} = ah + b.$$

1 Estimer a.

Ifølge boken skal egentlig

$$\dot{h} = -\sqrt{\frac{2gh}{b^2 - 1}}$$

der g er tyngdeakselerasjonen og b er forholdet mellom arealet til badestampen og arealet til hølet: `https://en.wikipedia.org/wiki/Torricelli%27s_law`

2 Stampen hadde radius på ca en meter, men jeg glemte å måle radien på hølet. Hva var den?

Figur 1: Elsys Torjus koser seg i badestampen

Figur 2: Elsys Torjus hopper i havet

Noen antikvariske bøker velger å sette opp diskret fourieromvending på en litt annen måte. De tar den vanlige fourierkoeffisientformelen

$$c_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} x(t) e^{-int} \, dt$$

og bruker den sammensatte trapesregelen

$$\int_a^b f(t) \, dt \approx \frac{1}{n} \left(\frac{f(a)}{2} + \sum_{k=1}^{n-1} f\left(a + k \frac{b-a}{n} \right) + \frac{f(b)}{2} \right)$$

på denne. Da får du en diskret fourieromvendingsformel som går på et litt annet gitter og likner noe mer på den klassiske fourierkoeffisientformelen.

3 Prøv.