

## 2 - 6 - HVA ER EN KATT?

Nå skal vi se litt på hvordan lineæralgebra brukes i datahåndtering; så og si alt vi har lært dette året kommer til nytte. Vi skal ikke bry oss med "statistikk" som sådan, for i statistikk er man veldig opptatt av å skille mellom parametrene i sannsynlighetsmodellen som ligger til grunn og estimatorene man stapper dataene inn i (for eksempel gjennomsnitt eller empirisk standardavvik). Vi skal bry oss mest med lineæralgebraknepene.

- 1 Da må vi først observere noe gunstig om matriser med ortogonale kolonner. Finn den inverse til matrisen

$$\begin{pmatrix} 1 & 1 & 1 \\ 0 & -\sqrt{2} & \sqrt{2} \\ -1 & 1 & 1 \end{pmatrix}$$

uten å gausseeliminere - se heller på oppgave 9 i økt 2-4.

Vi sier at en reell matrise  $P$  **ortogonal** dersom den har ortonormale kolonner, og da er enkelt og greit

$$P^{-1} = P^T.$$

Dersom en matrise  $n \times n$ -matrise  $A$  har  $n$  ortogonale egenvektorer sier vi derfor at  $A$  er **ortogonalt diagonaliserbar**, og skriver

$$A = P\Lambda P^T.$$

- 2 Finn den ortogonale diagonaliseringen til  $\begin{pmatrix} 1 & 2 & 2 \\ 2 & 6 & 2 \\ 2 & 2 & 6 \end{pmatrix}$ .

Vi sier at en reell matrise  $A$  er **symmetrisk** dersom  $A = A^T$ . I anvendelser er det veldig vanlig at det dukker opp symmetriske matriser. En reell matrise er ortogonalt diagonaliserbar hvis og bare hvis den er symmetrisk. Dette kalles **spektralteoremet**, og er ikke trivielt å vise, men det er heldigvis trivielt å huske.

- 3 Klarer du å ortogonalt diagonalisere  $\begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}$ ?



Som du kanskje ser av oppgaven over, kan en symmetrisk matrise fint ha en ikke-ortogonal egenvektorbasis, men poenget er at du alltid kan finne en som er ortogonal. Hvordan vi finner en ortogonal diagonalisering for matrisen i oppgave 3 kommer vi tilbake. Nå tar vi litt om datahåndtering. Det er en matrise som er veldig praktisk, nemlig kolonnevektoren

$$\mathbf{1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

Det er ikke helt uvanlig å skrive  $\mathbf{1}_n$  for å indikere dimensjon, men vanligvis er dette overflødig.

4 Regn ut  $\mathbf{1}^T \mathbf{1}$ ,  $\mathbf{1} \mathbf{1}^T$  og  $\mathbf{1}^T \mathbf{x}$ .

I faktiske anvendelser er det vanlig å **sentrere** datavektoren  $\mathbf{x}$  ved å lage en ny datavektor der man har trukket  $\bar{x}$  fra alle komponentene i  $\mathbf{x}$ , og så jobbe med denne istedet. Med den nye vektoren  $\mathbf{1}$  kan vi skrive den sentrerte variabelen som

$$\mathbf{x} - \bar{x} \cdot \mathbf{1} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} - \bar{x} \cdot \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} - \begin{pmatrix} \bar{x} \\ \bar{x} \\ \vdots \\ \bar{x} \end{pmatrix} = \begin{pmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \vdots \\ x_n - \bar{x} \end{pmatrix} = \left( I - \frac{1}{n} \mathbf{1} \mathbf{1}^T \right) \mathbf{x}$$

Matrisen

$$C_n = \left( I - \frac{1}{n} \mathbf{1} \mathbf{1}^T \right)$$

kalles **sentreringsmatrisen** fordi den sentrerer  $\mathbf{x}$ . Vi sier at en lineæroperator  $P$  er en **projeksjon** dersom  $P^2 = P$ . Dette betyr kort og godt at om du bruker lineæroperatoren to ganger etter hverandre, skjer det ingenting når du bruker den for andre gang.

5 Vis at  $C_n$  er en projeksjon og finn gjennomsnittet til den nye sentrerte variabelen.



Med sentrerte variable blir livet greit. I økt 2-6 utledet du uttrykket

$$\beta_1 = \frac{\mathbf{x}^T \mathbf{y} - n \bar{x} \bar{y}}{\mathbf{x}^T \mathbf{x} - n (\bar{x})^2} \quad \text{og} \quad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

for koeffisientene til det rettlinjede regresjonspolynommet.

- 6 Dette ser mye penere ut om  $\mathbf{x}$  og  $\mathbf{y}$  er sentrerte, og forklarer hvorfor noen sier at man “projiserer  $\mathbf{y}$  på  $\mathbf{x}$ ” når man kjører regresjon.

Skjønte du forrige side, kan du gå med på at sentrerte variable er lurt. La oss spinne videre litt. **Multipel regresjon** handler om at forklaringsvariabelen har flere dimensjoner. Du kan for eksempel tenke at du har målt sylindervolum  $\mathbf{x}_1$  og vekt  $\mathbf{x}_2$  på  $n$  biler, og så er responsvariabelen  $\mathbf{y}$  drivstofforbruk per kilometer. Hvis du ønsker å tilpasse et første ordens regresjonspolynom i to variable, får du det overbestemte likningssettet

$$\mathbf{y} = \beta_{10} \mathbf{x}_1 + \beta_{01} \mathbf{x}_2 + \beta_0 \mathbf{1}$$

eller

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \beta_{10} \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{pmatrix} + \beta_{01} \begin{pmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{n2} \end{pmatrix} + \beta_0 \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}$$

og dette kan angripes med minste kvadraters metode slik som i forrige uke.

- 7 Regn ut formlene for regresjonskoeffisientene dersom  $\mathbf{x}_1$ ,  $\mathbf{x}_2$  og  $\mathbf{y}$  er sentrerte.

La oss anta du har  $n$  realiseringer av en  $p$ -dimensjonal tilfeldig variabel. Det vanlig å sette disse opp i en  $n \times p$ -matrise:

$$X = \left( \begin{array}{c|c|c|c} \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_p \end{array} \right)$$

Tenk at du måler lengde, bredde og høyde på alle plankene i en plankestabel, og så sentrerer du hver av disse. Til slutt setter du alt opp i matrisen  $X$  på en slik måte at rad  $k$  er de sentrerte målene til planke  $k$ . Vi kaller  $\frac{1}{n-1} X^T X$  den **empiriske kovariansmatrisen**, og den er en estimator for den antatte bakenforliggende kovariansmatrisen.<sup>1</sup>

- 8 Vis at elementene i  $\frac{1}{n-1} X^T X$  er den empiriske kovariansen mellom kolonnene i  $X$ .



<sup>1</sup>[https://en.wikipedia.org/wiki/Estimation\\_of\\_covariance\\_matrices](https://en.wikipedia.org/wiki/Estimation_of_covariance_matrices)

Vi sier at en matrise  $A$  er **positivt definit** dersom

$$\mathbf{x}^T A \mathbf{x} > 0$$

for alle  $\mathbf{x}$ , og **positivt semidefinit** dersom

$$\mathbf{x}^T A \mathbf{x} \geq 0$$

for alle  $\mathbf{x}$ .

- 9 Vis at den empiriske kovariansmatrisen er positivt semidefinit.

I flervariabel statistikk finnes det noe som kalles **prinsipalkomponentanalyse**.<sup>2</sup> Dette er en teknikk for å lete opp akser i datamengden der samvariasjonen er sterk. Teknikken baserer seg på å projisere  $X$  inn på en retning  $\mathbf{v}$  slik at den empiriske variansen til de resulterende punktene blir maksimert. Dette får man til ved å lete opp egenvektorene til de største egenverdiene til kovariansmatrisen.

- 10 Forklar hvorfor egenvektorene til  $X^T X$  forteller om samvariasjonen mellom kolonnene i  $X$ . (Hint: Finn ut hva en "Rayleigh quotient" er.)<sup>3</sup>

I TMA4245 skal du også drive med liknende ting mot slutten av semesteret, se her:

<https://tma4245.math.ntnu.no/forventing-og-variens/korrelasjon/>

og her:

<https://tma4245.math.ntnu.no/forventing-og-variens/kovarians/>

- 11 Er det samvariasjon mellom høyde og bredde i kaibordene fra sommeren 2023?  
<https://folk.ntnu.no/mortano/python/kai/kai.csv>

Noen ganger er det regnet som ugunstig dersom kolonnene i datamatriksen har veldig forskjellig lengde. Dette fikser vi ved å dele ut standardavviket i hver kolonne, akkurat som når man normaliserer vektorer ellers i lineær algebra.



<sup>2</sup>[https://en.wikipedia.org/wiki/Principal\\_component\\_analysis](https://en.wikipedia.org/wiki/Principal_component_analysis)

<sup>3</sup>[https://en.wikipedia.org/wiki/Rayleigh\\_quotient](https://en.wikipedia.org/wiki/Rayleigh_quotient)

Nå skal vi se at det er faktisk ikke nødvendig å gange datamatriksen med seg selv for å finne kovariansmatrisen. Det finnes noe enda mer sofistikert som heter SVD. La oss begynne med litt geometrisk fikling.

12 Skriv en rutine som tar inn  $A \in \mathbb{R}^{2 \times 2}$  og plotter  $Ax$  for alle  $x$  på enhetssirkelen.

Fikk du til dette, vil du oppdage at koden din alltid produserer ellipser, ihvertfall så lenge  $A$  har linært uavhengige kolonner. La oss finne ut mer om disse ellipsoidene. Trikset er å studere  $A^T A$  og  $AA^T$ . Vi lar  $n > p$ , slik at  $A$  er høyt og tynn og  $A^T$  er lang og flat. Dette er ikke noen essensiell restriksjon, men vi gjør den fordi det hjelper litt på visualiseringen.

13 Både  $A^T A$  og  $AA^T$  er ortogonalt diagonaliserbare. Hvorfor?

Siden  $A^T A$  er ortogonalt diagonaliserbar, finnes det et ortonormalt sett med egenvektorer. La oss sette disse opp i en ortogonal matrise  $V$ . Det samme gjelder for  $AA^T$ ; la oss sette disse egenvektoren opp i en ortonormal matrise  $U$ .

14 Hva er dimensjonene til  $A^T A$ ,  $AA^T$ ,  $V$  og  $U$ ?  
Hvor mange av egenverdiene til  $AA^T$  kan være forskjellige fra null?

Her er enda en observasjon.

15 Vis at dersom  $v$  er en egenvektor til  $A^T A$  med egenverdi  $\lambda$ , er  $Av$  en egenvektor til  $AA^T$ . Hva er egenverdien? Kan du si noe om fortegnet?

Hvis du skjønnte forrige oppgave, skjønner du forhåpentligvis at  $A^T A$  og  $AA^T$  har de samme egenverdiene så lenge du ser bort fra multiplisiteten til de egenverdiene som er null. La oss nå anta at  $v$  er en egenvektor til  $A$  med lengde 1.

16 Hva er lengden til  $Av$ ?

Lengden til  $Av$  kalles **singulærverdi**, og vi bruker bokstaven  $\sigma_k$ . Kolonnene i  $U$  og  $V$  kalles henholdsvis de **venstre og høyre singulærvektorene**. Hvis du tenker nøye på alt du har gjort til nå, vil du se at det går an å skrive

$$A = U\Sigma V^T$$

der  $\sigma$ -ene sitter i den rektangulære diagonalmatrisen  $\Sigma$ . Dette kalles **SVD-faktoriseringen**<sup>4</sup> til  $A$ . Denne finnes alltid, og dersom du sorterer singulærverdiene i synkende rekkefølge ( $\sigma_1$  er størst) og velger alle positive, er  $\Sigma$  entydig. De unitære matrisene  $U$  og  $V$  vil ikke nødvendigvis være entydige, men de er det i noen tilfeller.

17 Hvordan ser  $\Sigma$  ut? Finn svd-faktoriseringen til

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

Singulærverdidekomposisjonen er en av de viktigste matrisefaktoriseringene. Har du den har du mye, og pythonkommandoen heter `numpy.linalg.svd`. Nå skal vi se litt på hva svd kan brukes til. Det første

<sup>4</sup>[https://en.wikipedia.org/wiki/Singular\\_value\\_decomposition](https://en.wikipedia.org/wiki/Singular_value_decomposition)

er kompresjon. Vi kan skrive  $A$  slik:

$$A = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^* + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^* + \sigma_3 \mathbf{u}_3 \mathbf{v}_3^* + \dots + \sigma_r \mathbf{u}_r \mathbf{v}_r^*$$

der  $r$  er antallet singularverdier som ikke er null. Vi ser nå at dersom det er bare er to av singularverdiene som er over for eksempel maskinpresisjon, er all informasjon i matrisen inneholdt i uttrykket

$$A = \mathbf{u}_1 \sigma_1 \mathbf{v}_1^* + \mathbf{u}_2 \sigma_2 \mathbf{v}_2^*.$$

- 18] La oss lage et bilde av en katt. Lag deg en matrise  $A$  i python med tilfeldige tall mellom -1 og 1. Dette er bakgrunnen. Sett så ett element til 10 eller 100 eller noe sånt. Dette er katten. Kjør svd på  $A$  og skriv ut

$$A - \mathbf{u}_1 \sigma_1 \mathbf{v}_1^* \quad \text{og} \quad A - (\mathbf{u}_1 \sigma_1 \mathbf{v}_1^* + \mathbf{u}_2 \sigma_2 \mathbf{v}_2^*).$$

## UKENS NØTTER

Vi sier at lengden

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

er **indusert** av indreproduktet  $(\cdot, \cdot)$ . Men det finnes også lengder som ikke er indusert av et indreprodukt. Derfor har vi egne aksiomer for lengde:

- $\|\mathbf{x}\| > 0$  dersom  $\mathbf{x} \neq \mathbf{0}$
- $\|a\mathbf{x}\| = |a|\|\mathbf{x}\|$
- $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$

På fagspråket sier man **norm** istedet for lengde. Det finnes mange normer som ikke er indusert av indreprodukt

- 1] Vis at  $\|\mathbf{x}\|_2 = \sqrt{(\mathbf{x}, \mathbf{x})}$  er en lengde.

- 2] Vis at  $\|\mathbf{x}\|_\infty = \max_k |x_k|$  og  $\|\mathbf{x}\|_1 = \sum_k |x_k|$  er lengder når  $\mathbf{x} \in \mathbb{C}^n$ . Forklar subskriptene.

Så har vi **matrisenormen**  $\|A\| = \max_{\|\mathbf{x}\|=1} \|A\mathbf{x}\|$ .

- 3] Vis at dette er en lengde.

- 4] Vis at  $e^A$  er absolutt konvergent for alle  $A$ .

