Tillämpningar av SVD

TATA53 LÖK

Jonathan Nilsson

Linköping Universitet

Del I

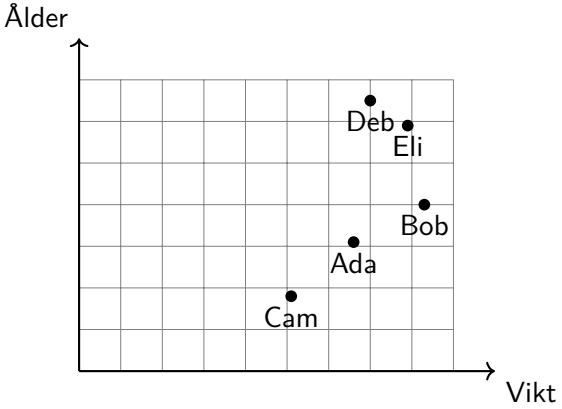
Data-analys med SVD

Cosinus-similaritet

$$\text{cosSim}(u, v) := \frac{u \bullet v}{\|u\| \cdot \|v\|}.$$

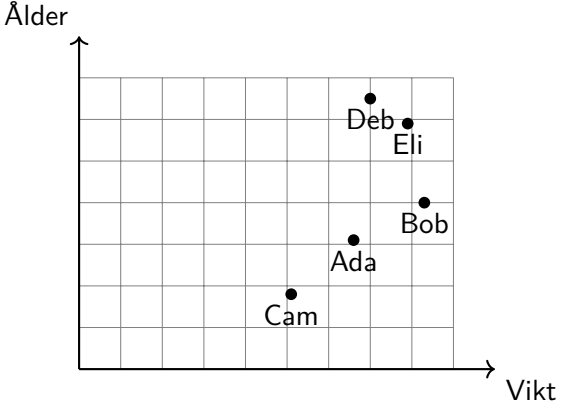
Vikt och ålder för Ada, Bob, Cam, Deb, Eli:

$$A = \begin{pmatrix} 66 & 31 \\ 83 & 40 \\ 51 & 18 \\ 70 & 65 \\ 79 & 59 \end{pmatrix}$$



Vikt och ålder för Ada, Bob, Cam, Deb, Eli:

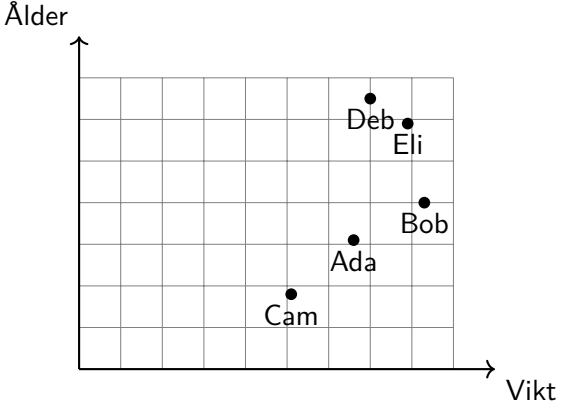
$$A = \begin{pmatrix} 66 & 31 \\ 83 & 40 \\ 51 & 18 \\ 70 & 65 \\ 79 & 59 \end{pmatrix}$$



$$\text{cosSim}(A_1, A_2) = 0.955$$

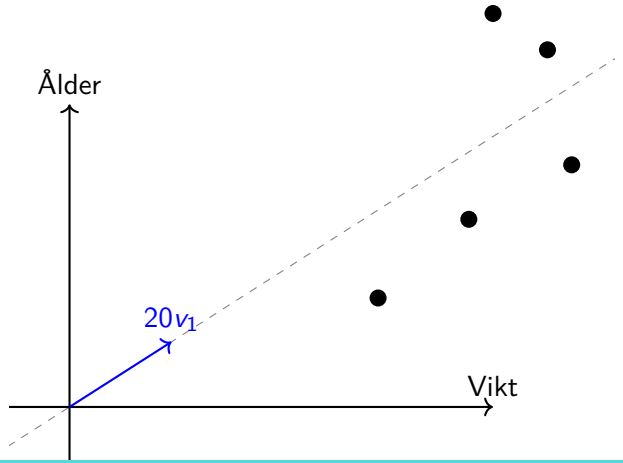
Vikt och ålder för Ada, Bob, Cam, Deb, Eli:

$$A = \begin{pmatrix} 66 & 31 \\ 83 & 40 \\ 51 & 18 \\ 70 & 65 \\ 79 & 59 \end{pmatrix}$$



$$\text{cosSim}(A_1, A_2) = 0.955 \qquad \text{cosSim}(\text{Bob}, \text{Deb}) = 0.953$$

Riktning med högst variation



Slutsats

De första singulärvektorerna pekar i riktningarna där data varierar mest. Genom att projicera data på rummet som spänns upp av de första singulärvektorerna reducerar vi dimensionen på data så att den blir så utspriden som möjligt.

Filmbetyg

Sex personer har betygsatt fem filmer:

	Interstellar	Mean Girls	The Matrix	Borat	Ghost Busters
Alex	5	2	4	2	4
Bert	5	1	4	1	4
Cleo	4	5	2	4	5
Dany	3	5	3	5	5
Elle	5	2	5	2	3
Fred	5	3	4	2	4

Filmbetyg

Sex personer har betygsatt fem filmer:

	Interstellar	**Mean Girls**	**The Matrix**	**Borat**	**Ghost Busters**
Alex	5	2	4	2	4
Bert	5	1	4	1	4
Cleo	4	5	2	4	5
Dany	3	5	3	5	5
Elle	5	2	5	2	3
Fred	5	3	4	2	4

Låt A vara 6×5-matrisen med dessa koefficienter.

SVD av film-matrisen

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

$$U = \begin{pmatrix} 0.39 & 0.26 & 0.12 & 0.17 & -0.46 \\ 0.36 & 0.45 & 0.41 & 0.52 & 0.15 \\ 0.44 & -0.49 & 0.48 & -0.38 & -0.35 \\ 0.45 & -0.57 & -0.42 & 0.49 & 0.22 \\ 0.39 & 0.37 & -0.64 & -0.30 & -0.29 \\ 0.41 & 0.16 & 0.09 & -0.46 & 0.71 \end{pmatrix} \quad \Sigma = \begin{pmatrix} 20.1 & & & & \\ & 5.60 & & & \\ & & 1.57 & & \\ & & & 0.96 & \\ & & & & 0.35 \end{pmatrix} \quad V = \begin{pmatrix} 0.54 & 0.45 & 0.33 & -0.42 & -0.47 \\ 0.38 & -0.56 & -0.05 & -0.61 & 0.41 \\ 0.44 & 0.47 & -0.67 & 0.14 & 0.36 \\ 0.34 & -0.50 & -0.41 & 0.28 & -0.63 \\ 0.51 & -0.13 & 0.53 & 0.60 & 0.30 \end{pmatrix}$$

Singulärvärden motsvarar “latenta faktorer”

$$\sigma_1 = 20.1 \sim \text{Filmers populäritet}$$

Singulärvärden motsvarar “latenta faktorer”

$$\sigma_1 = 20.1 \sim \text{Filmers populäritet}$$

$$\sigma_2 = 5.60 \sim \text{Filmers genre}$$

Högersingulärvektorer i film-rummet

I film-rummet $\mathbb{R}^5$ motsvarar varje film en basvektor:

(Interstellar,Mean Girls, The Matrix, Borat, Ghost Busters)

Högersingulärvektorer i film-rummet

I film-rummet $\mathbb{R}^5$ motsvarar varje film en basvektor:

(Interstellar,Mean Girls, The Matrix, Borat, Ghost Busters)

$v_1 = (0.54, 0.38, 0.44, 0.34, 0.51)^T \sim$ Populäritet för varje film

Högersingulärvektorer i film-rummet

I film-rummet $\mathbb{R}^5$ motsvarar varje film en basvektor:

(Interstellar,Mean Girls, The Matrix, Borat, Ghost Busters)

$$v_1 = (0.54, 0.38, 0.44, 0.34, 0.51)^T \sim \text{Populäritet för varje film}$$

$$v_2 = (0.45, -0.56, 0.47, -0.50, -0.13)^T \sim \text{Filmgenre för varje film}$$

Vänstersingulärvektorer i tittar-rummet

I tittar-rummet $\mathbb{R}^6$ motsvarar varje tittare en basvektor:

(Alex, Burt, Cleo, Dany, Elle, Fred)

Vänstersingulärvektorer i tittar-rummet

I tittar-rummet $\mathbb{R}^6$ motsvarar varje tittare en basvektor:

$$\text{(Alex, Burt, Cleo, Dany, Elle, Fred)}$$

$$u_1 = (0.39, 0.36, 0.44, 0.45, 0.39, 0.41)^T \sim \text{Genomsnittsbetyg för varje tittare}$$

Vänstersingulärvektorer i tittar-rummet

I tittar-rummet $\mathbb{R}^6$ motsvarar varje tittare en basvektor:

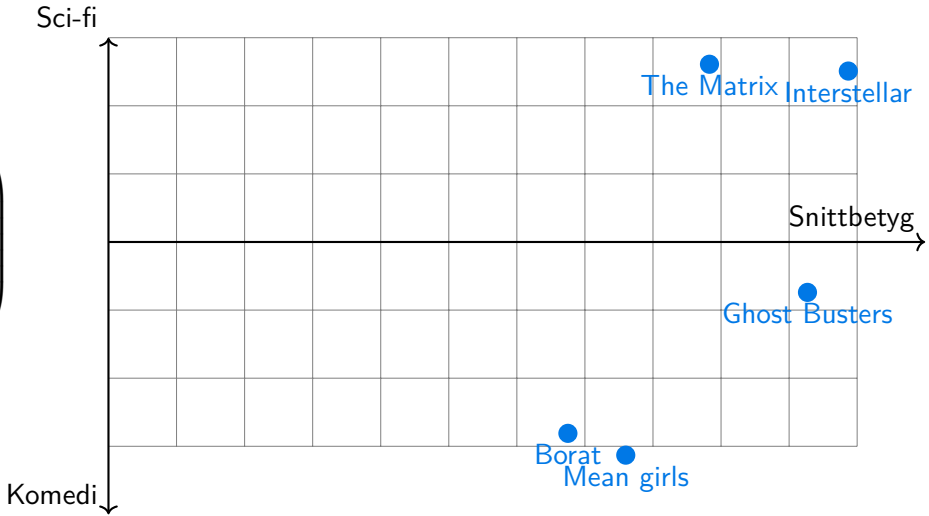
(Alex, Burt, Cleo, Dany, Elle, Fred)

$$u_1 = (0.39, 0.36, 0.44, 0.45, 0.39, 0.41)^T \sim \text{Genomsnittsbetyg för varje tittare}$$

$$u_2 = (0.26, 0.45, -0.49, -0.57, 0.37, 0.1)^T \sim \text{Genre-preferens för varje tittare}$$

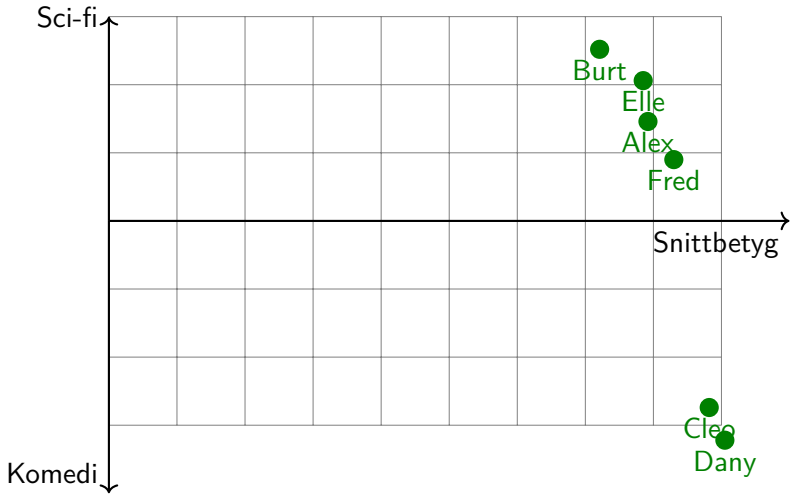
Filmer i tittar-rummet

$$\begin{pmatrix} 10.87 & 2.51 \\ 7.60 & -3.13 \\ 8.83 & 2.61 \\ 6.75 & -2.81 \\ 10.27 & -0.74 \end{pmatrix}$$



Tittare i film-rummet

$$\begin{pmatrix} 7.92 & 1.46 \\ 7.21 & 2.52 \\ 8.82 & -2.74 \\ 9.05 & -3.22 \\ 7.85 & 2.06 \\ 8.30 & 0.90 \end{pmatrix}$$



Del II

Ansiktsigenkänning

Ansikten från MAI

Input: 70 ansikten av anställda vid MAI på LiU.

Ansikten från MAI

Input: 70 ansikten av anställda vid MAI på LiU.

Original → Justerade → Omskalade svartvita → Snitt-subtraherade

Snittansiktet

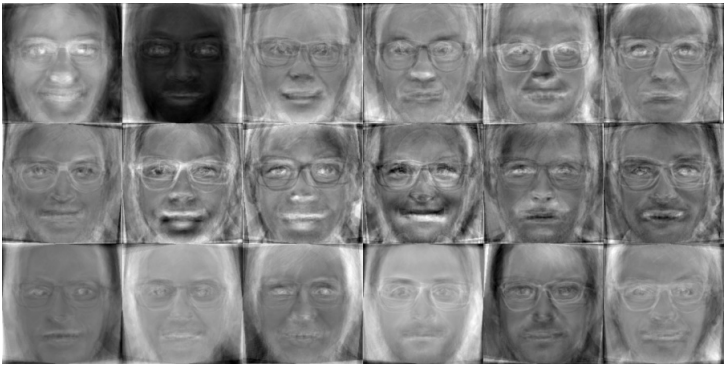
$$c = \frac{1}{70} \sum_{i=1}^{70} z_i =$$

Singulärvärdesfaktorisering

A är nu en 70×22500-matris. Vi tar fram egenvärden till $A^T A$.

Singulärvärdesfaktorisering

A är nu en 70×22500-matris. Vi tar fram egenvärden till $A^T A$.



De första 18 *egenansiktena* (egenvektorer till $A^T A$).

Projektion på principalkomponenterna

$k = 10$

Projektion på principalkomponenterna

$k = 10$

$k = 25$

Projektion på principalkomponenterna

$k = 10$

$k = 25$

$k = 40$

Projektion på principalkomponenterna

$k = 10$

$k = 25$

$k = 40$

$k = 55$

Projektion på principalkomponenterna

$k = 10$

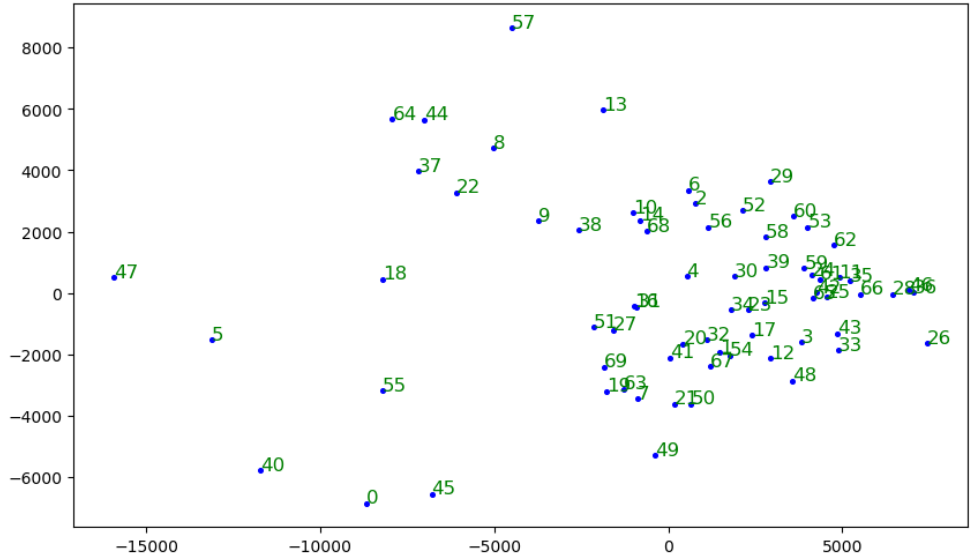
$k = 25$

$k = 40$

$k = 55$

$k = 70$

Dimensionsreduktion av ansiktsrummet



Del III

Släktskapsanalys

Genfrekvenser

En gen är en kort sekvens DNA. Vi väljer 35 olika gener och går igenom DNA för ett antal djur och räknar hur många gånger varje sådan gen förekommer i dess genom.

Genfrekvenser

En gen är en kort sekvens DNA. Vi väljer 35 olika gener och går igenom DNA för ett antal djur och räknar hur många gånger varje sådan gen förekommer i dess genom.

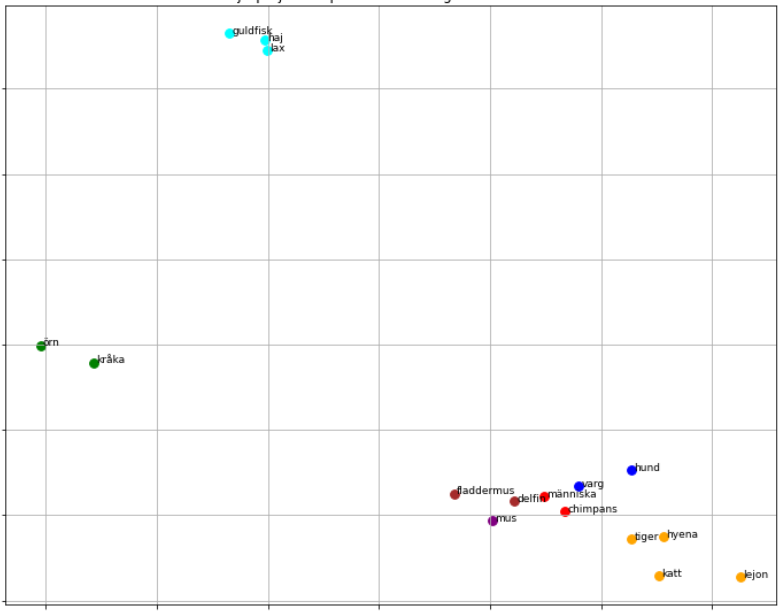
	Gen																																
Djur	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
människa	16	16	14	15	14	16	14	14	19	13	23	23	24	5	10	2	5	4	11	5	4	9	4	6	8	3	4	6	8	10	4	2	6	7	9
chimpans	13	16	19	17	22	18	14	11	16	12	27	23	25	6	2	6	5	5	5	8	1	3	4	11	9	4	8	7	5	5	5	9	3	3	9
mus	15	19	17	17	21	14	18	22	16	18	5	8	9	4	3	6	2	2	5	17	17	5	6	4	4	8	5	4	3	6	9	6	4	4	6
katt	20	16	16	18	18	17	15	15	18	17	6	11	8	22	28	24	4	5	7	3	5	6	3	4	2	4	7	2	2	8	7	6	7	3	6
hund	18	13	15	18	13	15	19	15	19	16	8	6	3	5	4	7	30	24	21	7	9	5	7	6	6	9	5	10	12	7	1	8	8	5	7
varg	14	17	14	17	15	22	15	15	15	14	5	7	4	3	4	4	29	26	25	10	2	9	11	7	8	6	5	6	3	8	4	5	3	8	6
lejon	21	17	20	19	16	15	15	20	11	17	6	5	6	29	27	27	5	4	7	7	3	3	6	9	6	8	8	5	5	9	8	7	7	2	7
tiger	15	13	15	14	18	18	13	14	17	14	11	5	6	24	23	27	4	1	6	7	9	6	10	8	5	11	2	7	5	5	6	6	7	9	6
hyena	13	19	17	11	19	14	17	14	15	15	4	5	6	27	25	25	5	6	5	8	4	7	9	10	5	10	3	6	10	5	6	8	7	2	11
fladdermus	17	15	16	16	12	15	17	16	17	19	6	8	3	6	2	1	2	6	8	7	3	15	18	15	6	7	2	4	8	9	7	4	5	7	5
delfin	12	13	20	17	14	16	19	14	19	19	10	8	5	6	4	4	10	4	4	7	3	20	20	19	5	4	0	4	2	4	8	5	8	6	4
örn	9	6	6	4	2	6	9	3	2	4	7	4	4	8	5	8	4	6	6	3	7	8	11	8	3	6	18	21	24	20	2	8	9	5	7
kråka	7	2	5	2	9	10	6	6	13	10	5	8	7	7	5	4	6	4	7	4	6	5	5	3	4	10	23	17	22	17	6	6	3	3	11
haj	7	8	7	3	4	4	6	5	11	10	5	1	5	4	4	7	9	4	6	6	5	8	6	4	7	5	2	2	10	10	27	26	26	26	28
lax	4	3	6	7	7	2	6	8	11	8	9	5	4	10	5	2	9	8	6	3	3	8	4	3	4	6	5	7	10	8	27	25	25	26	26
guldfisk	3	7	4	7	6	7	4	3	4	6	5	5	12	6	4	6	6	9	9	5	7	10	2	5	5	5	7	4	7	4	28	25	27	25	28

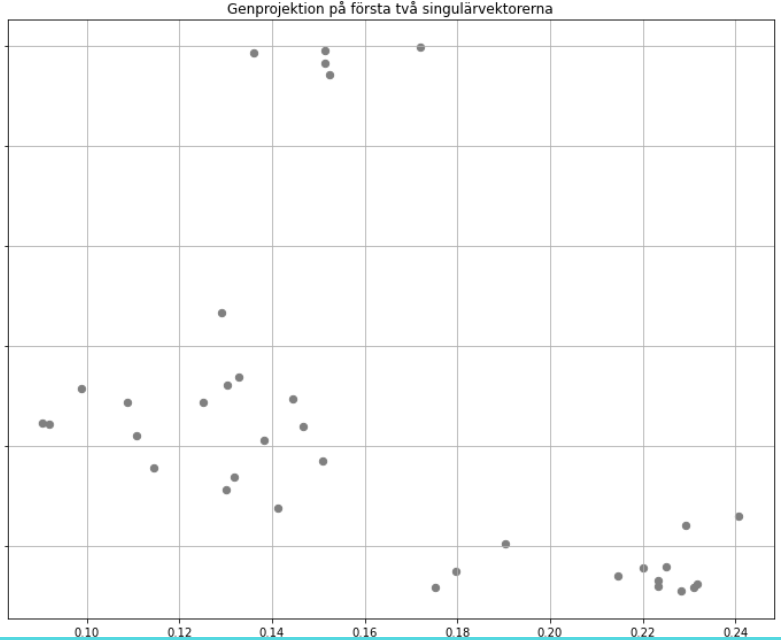
Genfrekvenser

En gen är en kort sekvens DNA. Vi väljer 35 olika gener och går igenom DNA för ett antal djur och räknar hur många gånger varje sådan gen förekommer i dess genom.

	Gen																																
Djur	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
människa	16	16	14	15	14	16	14	14	19	13	23	23	24	5	10	2	5	4	11	5	4	9	4	6	8	3	4	6	8	10	4	2	6	7	9
chimpans	13	16	19	17	22	18	14	11	16	12	27	23	25	6	2	6	5	5	5	8	1	3	4	11	9	4	8	7	5	5	5	9	3	3	9
mus	15	19	17	17	21	14	18	22	16	18	5	8	9	4	3	6	2	2	5	17	17	5	6	4	4	8	5	4	3	6	9	6	4	4	6
katt	20	16	16	18	18	17	15	15	18	17	6	11	8	22	28	24	4	5	7	3	5	6	3	4	2	4	7	2	2	8	7	6	7	3	6
hund	18	13	15	18	13	15	19	15	19	16	8	6	3	5	4	7	30	24	21	7	9	5	7	6	6	9	5	10	12	7	1	8	8	5	7
varg	14	17	14	17	15	22	15	15	15	14	5	7	4	3	4	4	29	26	25	10	2	9	11	7	8	6	5	6	3	8	4	5	3	8	6
lejon	21	17	20	19	16	15	15	20	11	17	6	5	6	29	27	27	5	4	7	7	3	3	6	9	6	8	8	5	5	9	8	7	7	2	7
tiger	15	13	15	14	18	18	13	14	17	14	11	5	6	24	23	27	4	1	6	7	9	6	10	8	5	11	2	7	5	5	6	6	7	9	6
hyena	13	19	17	11	19	14	17	14	15	15	4	5	6	27	25	25	5	6	5	8	4	7	9	10	5	10	3	6	10	5	6	8	7	2	11
fladdermus	17	15	16	16	12	15	17	16	17	19	6	8	3	6	2	1	2	6	8	7	3	15	18	15	6	7	2	4	8	9	7	4	5	7	5
delfin	12	13	20	17	14	16	19	14	19	19	10	8	5	6	4	4	10	4	4	7	3	20	20	19	5	4	0	4	2	4	8	5	8	6	4
örn	9	6	6	4	2	6	9	3	2	4	7	4	4	8	5	8	4	6	6	3	7	8	11	8	3	6	18	21	24	20	2	8	9	5	7
kråka	7	2	5	2	9	10	6	6	13	10	5	8	7	7	5	4	6	4	7	4	6	5	5	3	4	10	23	17	22	17	6	6	3	3	11
haj	7	8	7	3	4	4	6	5	11	10	5	1	5	4	4	7	9	4	6	6	5	8	6	4	7	5	2	2	10	10	27	26	26	26	28
lax	4	3	6	7	7	2	6	8	11	8	9	5	4	10	5	2	9	8	6	3	3	8	4	3	4	6	5	7	10	8	27	25	25	26	26
guldfisk	3	7	4	7	6	7	4	3	4	6	5	5	12	6	4	6	6	9	9	5	7	10	2	5	5	5	7	4	7	4	28	25	27	25	28

Vi genomför SVD av genfrekvensmatrisen och reducerar till dimension 2 genom att projicera data på de första två singulärvektorerna.


Djurprojektion på första två singulärvektorerna
guldfisk
haj
lax
örn
kråka
fladdermus
delfin
människa
varg
hund
chimpans
mus
tiger
hyena
katt
lejon
40
30
20
10
0
−10
−20
40
45
50
55
60
65
70


Genprojektion på första två singulärvektorerna
0.4
0.3
0.2
0.1
0.0
−0.1
0.10
0.12
0.14
0.16
0.18
0.20
0.22
0.24

Del IV

Totalminsta kvadratmetoden

Att anpassa ett k-plan till ett antal punkter

Totalminsta kvadratproblemet

Givet $x_i, \ldots, x_m \in \mathbb{R}^n$, vilket **affint delrum** S av dimension k ligger närmast punkterna i meningen att

$$\sum_{i=1}^{m} d(x_i, S)^2 \text{ är minimalt,}$$

där $d(x_i, S) = \min_{s \in S} \|x_i - S\|$.

Svaret

För att anpassa ett k-plan till punkter $x_1, \ldots, x_m \in \mathbb{R}^n$:

Svaret

För att anpassa ett k-plan till punkter $x_1, \ldots, x_m \in \mathbb{R}^n$:

- Låt $c = \frac{1}{m} \sum_{i=1}^{m} x_i$ vara punkternas medelvärde.

Svaret

För att anpassa ett k-plan till punkter $x_1, \ldots, x_m \in \mathbb{R}^n$:

- Låt $c = \frac{1}{m} \sum_{i=1}^{m} x_i$ vara punkternas medelvärde.
- Bilda en $m \times n$-matris A där kolonnerna är de normaliserade datapunkterna $x_i - c$.

Svaret

För att anpassa ett k-plan till punkter $x_1, \ldots, x_m \in \mathbb{R}^n$:

- Låt $c = \frac{1}{m} \sum_{i=1}^{m} x_i$ vara punkternas medelvärde.
- Bilda en $m \times n$-matris A där kolonnerna är de normaliserade datapunkterna $x_i - c$.
- Ta fram SVD för A, låt $v_1, \ldots, v_r$ vara högersingulärvektorerna.

Svaret

För att anpassa ett k-plan till punkter $x_1, \ldots, x_m \in \mathbb{R}^n$:

- Låt $c = \frac{1}{m} \sum_{i=1}^{m} x_i$ vara punkternas medelvärde.
- Bilda en $m \times n$-matris A där kolonnerna är de normaliserade datapunkterna $x_i - c$.
- Ta fram SVD för A, låt $v_1, \ldots, v_r$ vara högersingulärvektorerna.
- Det k-dimensionella affina delrummet som bäst approximerar punkterna i totalminsta kvadratmening är

$$S = c + \operatorname{span}(v_1, \ldots, v_k)$$

för varje $k = 1, 2, 3, \ldots$.

Vilka punkter i $\mathbb{R}^3$ ligger närmast $(1,5)$, $(1,2)$, $(4,2)$?

Vilka punkter i $\mathbb{R}^3$ ligger närmast $(1,5),\ (1,2),\ (4,2)$?

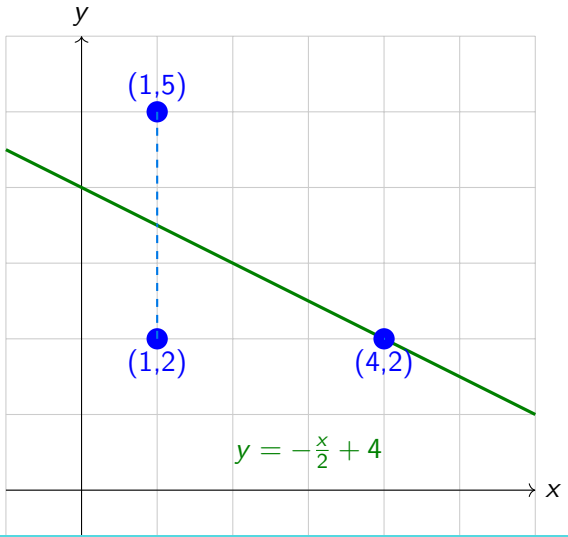
Gamla minsta kvadratmetoden:
Ansats $y = kx + m \ \rightarrow$ linjärt ekvationssystem $\rightarrow$ minstakvadratlösning.

Vilka punkter i $\mathbb{R}^3$ ligger närmast $(1,5)$, $(1,2)$, $(4,2)$?

Gamla minsta kvadratmetoden:
Ansats $y = kx + m$ $\rightarrow$ linjärt ekvationssystem $\rightarrow$ minstakvadratlösning.

Totalminsta kvadratmetoden:
Centrera punkterna kring origo $\rightarrow$ SVD av data-matrisen $\rightarrow$ singularvektorer ger riktningar för affina delrummet.

Vanliga minsta kvadratmetoden



Totalminsta kvadratmetoden