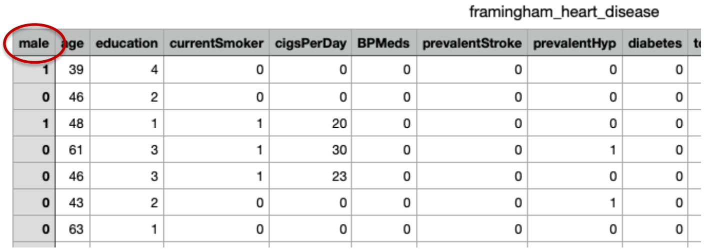
Assignment 5 Feedback

Applied Linear Algebra for Data Science

First…

Any comments on this?



framingham_heart_disease

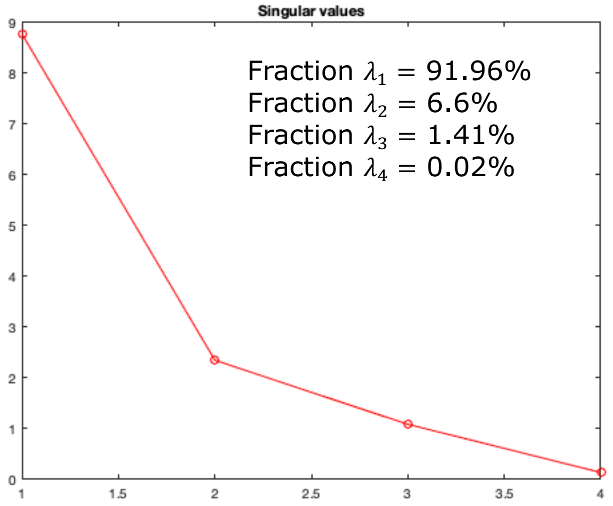
male	age	education	currentSmoker	cigsPerDay	BPMeds	prevalentStroke	prevalentHyp	diabetes	t
1	39	4	0	0	0	0	0	0	
0	46	2	0	0	0	0	0	0	
1	48	1	1	20	0	0	0	0	
0	61	3	1	30	0	0	1	0	
0	46	3	1	23	0	0	0	0	
0	43	2	0	0	0	0	1	0	
0	63	1	0	0	0	0	0	0	

Apparently there exist only male and non-males (=0) in the world

General things

- No need to form C explicitly $(C = \frac{1}{n-1}A^T A)$ and therefore relation between SVD of data A and eigenvalues/vectors of C
- …but A must be centered first!
- …and remove NaN's
- Principal components = AV or $U\Sigma$
- Note…
  "In both exercises below, use linear algebra built-in functions in Python in your code, such as built-in functions for SVD. Do not use higher-level libraries for PCA"

Q1a)



Fraction λ_1 = 91.96%
Fraction λ_2 = 6.6%
Fraction λ_3 = 1.41%
Fraction λ_4 = 0.02%

Almost all variance in the first two directions – reduce dimension to 2

Q1b)

- Variables = columns (movies) => work with $C = \frac{1}{n-1} A^T A$ and $n = 5$ (number of samples)
- First principal component ($= Av_1$ or $\sigma_1 u_1$):

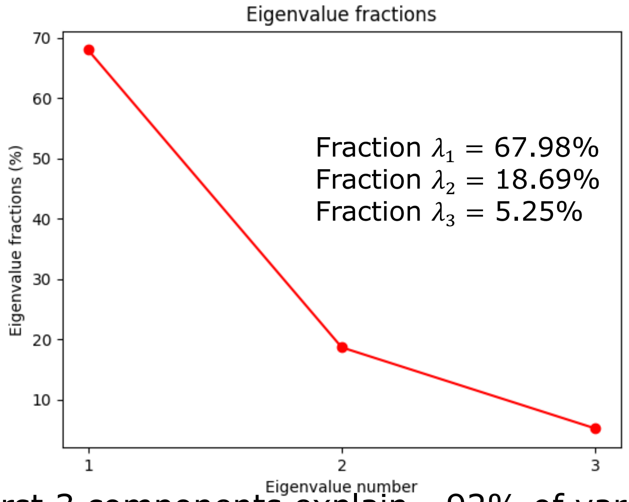
$$\begin{pmatrix} -3.9465 \\ 4.5370 \\ -1.7661 \\ -3.7083 \\ 4.8838 \end{pmatrix}$$

Look for ”orthogonal” groups

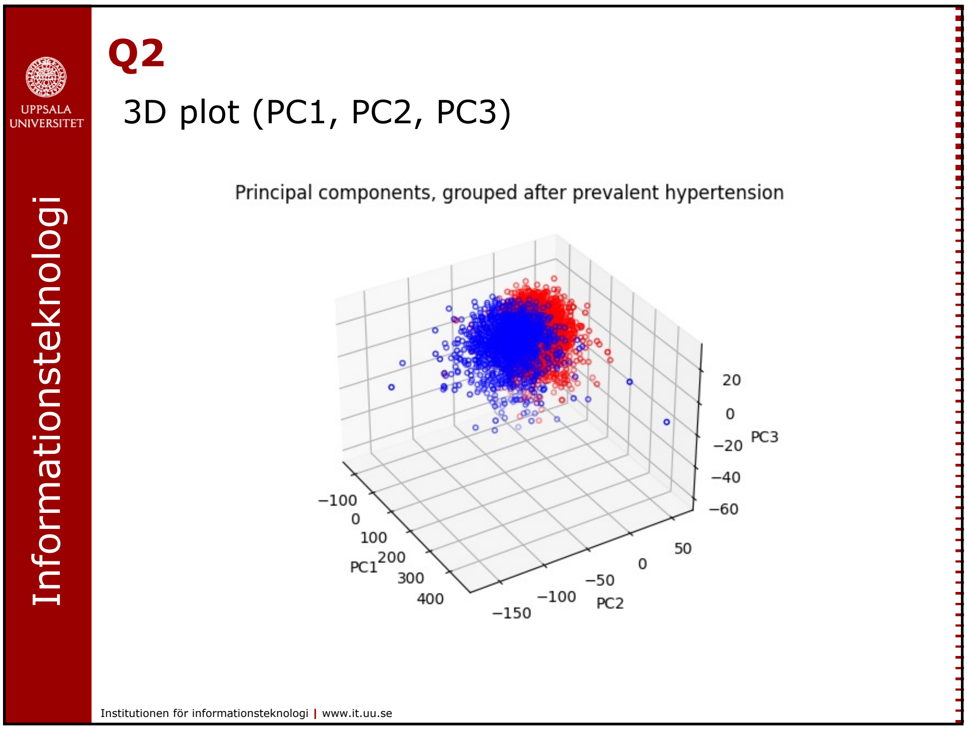
- The principle components show where we have the largest variance in the samples (explains 91.96% of the variance)
- Largest variance between Ali, Elsa, Johan on one side and Beatrix, Chandra on the other

Q2

- Dominating principal components:



- First 3 components explain ~92% of variance – can reduce dimension to 3



UPPSALA UNIVERSITET

Informationsteknologi

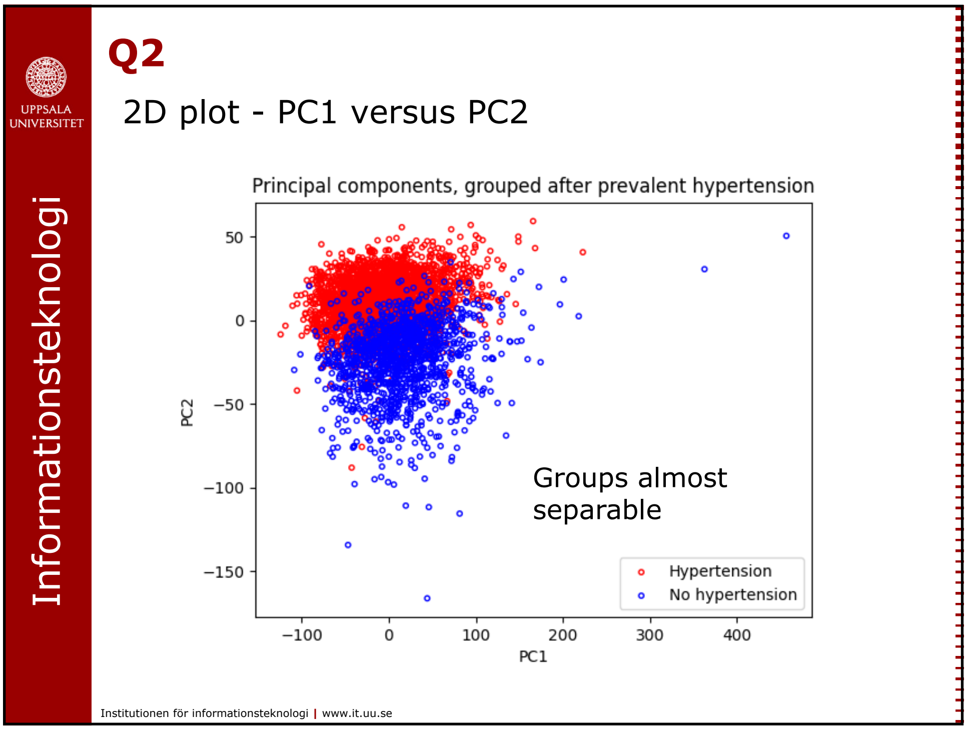
Q2

3D plot (PC1, PC2, PC3)

Principal components, grouped after prevalent hypertension

Institutionen för informationsteknologi | www.it.uu.se

8



UPPSALA UNIVERSITET

Informationsteknologi

Q2

2D plot - PC1 versus PC2

Principal components, grouped after prevalent hypertension

Groups almost separable

Institutionen för informationsteknologi | www.it.uu.se

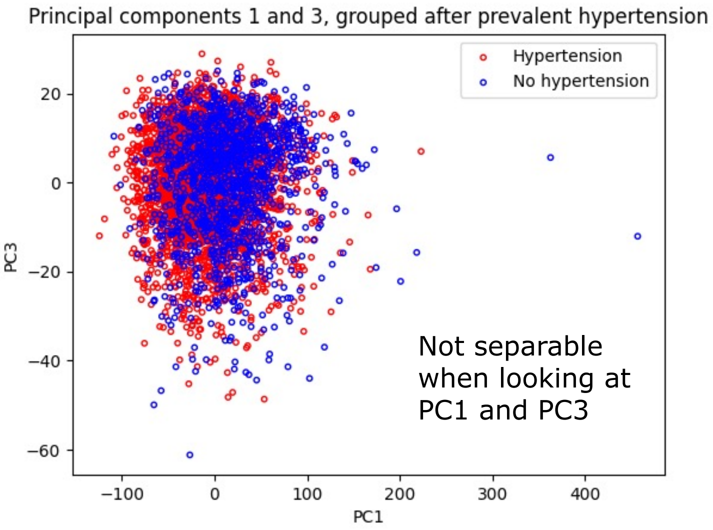
9

UPPSALA UNIVERSITET

Informationsteknologi

Q2

2D plot - PC1 versus PC3



Not separable when looking at PC1 and PC3

UPPSALA UNIVERSITET

Informationsteknologi

Q2

First eigenvector v_1

$$v_1 = \begin{array}{cl} \begin{pmatrix} 0.0571 \\ -0.01002 \\ 0.98541 \\ 0.14371 \\ 0.06253 \\ 0.01324 \\ 0.02938 \end{pmatrix} & \begin{array}{l} \text{Age} \\ \text{CigsPerDay} \\ \text{totChol} \\ \text{sysBP} \\ \text{diaBP} \\ \text{BMI} \\ \text{Heartrate} \end{array} \end{array}$$