MS0240
Week 6

Practicals 5b + 5c + 5d

Practical 5b

PCA on Covariance Matrix

PCA on Covariance Matrix

Example

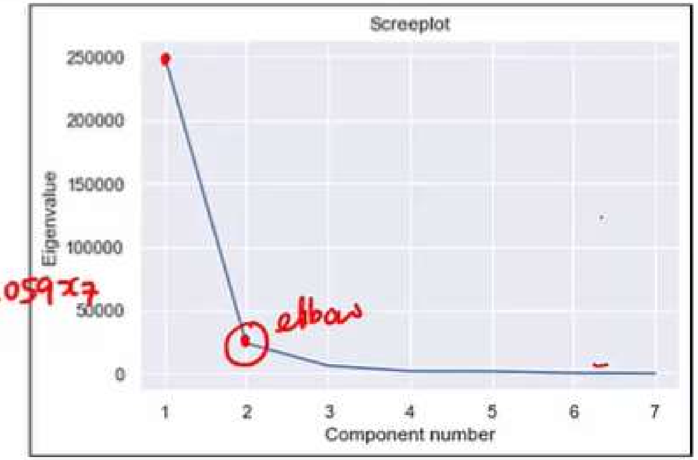
(a) PCA on covariance matrix

	Eigenvalue	Explained Variance	Cumulative Explained Variance	X_1 Bread	X_2 Vegetables	X_3 Fruit	X_4 Meat	X_5 Poultry	X_6 Milk	X_7 Wine
PC 1	250627.4594 λ_1	0.8795	0.8795	0.0685	0.3273	0.3038	0.7555	0.4621	0.0899	-0.0587
PC 2	23978.4709 λ_2	0.0841	0.9637	0.5477	0.4201	-0.0886	-0.0894	-0.2810	0.6397	0.1399
PC 3	6164.2178	0.0216	0.9853	0.4409	-0.3106	-0.3135	0.0558	0.3876	-0.1814	0.6516
PC 4	1952.6453	0.0069	0.9921	-0.0917	0.6924	0.2342	-0.3624	0.0919	-0.4330	0.3606
PC 5	1800.0528	0.0063	0.9985	-0.1745	-0.3209	0.6954	0.0386	-0.2490	0.2406	0.5115
PC 6	348.8297	0.0012	0.9997	0.6781	-0.1663	0.4698	-0.1162	-0.0744	-0.3759	-0.3624
PC 7	88.3866	0.0003	1.0000	-0.0479	-0.0991	0.2060	-0.5212	0.6937	0.4039	-0.1714

Total : 284960.0625

Number of PCs to extract:

- Kaiser's rule cannot be applied on covariance matrix.
- 1st PC already accounted for 88% of total variance.
- Scree plot shows elbow at PC2; suggesting 1 PC to extract.



Let's extract the first PC only.

PC: $\hat{y}=0.069x_1+0.327x_2+0.304x_3+0.756x_4+0.462x_5+0.090x_6-0.059x_7$

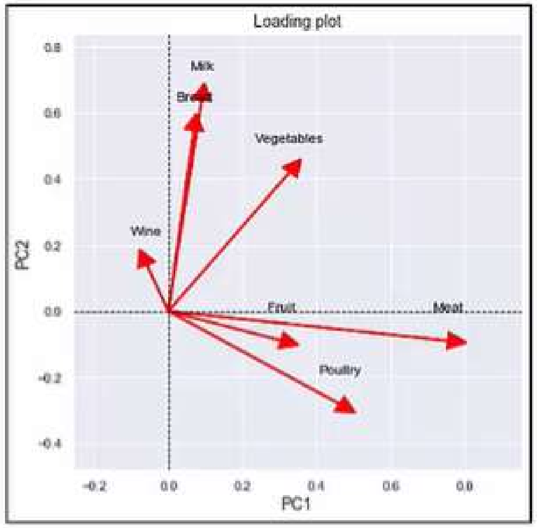
PCA on Covariance Matrix

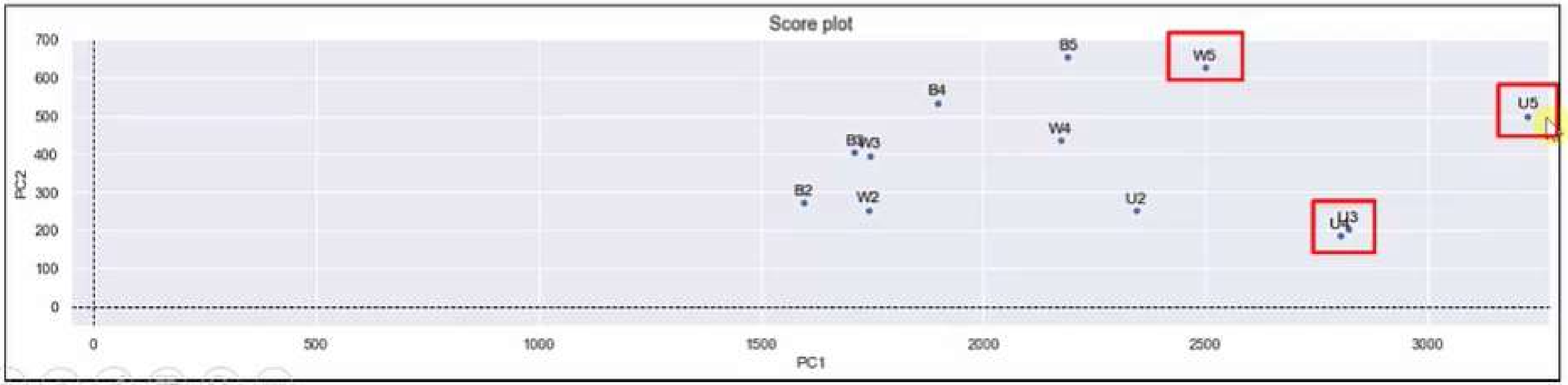
	Bread	Vegetables	Fruit	Meat	Poultry	Milk	Wine

PC: $\hat{y} = 0.069x_1 + 0.327x_2 + 0.304x_3 + 0.756x_4 + 0.462x_5 + 0.090x_6 - 0.059x_7$

The loadings on bread, milk and wine are quite small, whereas the loadings on vegetables, fruits, meat and poultry are bigger. The loading on wine is opposite in sign to the other loadings.

This PC seems to measure expenditure on food required of a balanced, and thus more affluent diet. (Note: wine was cheaper than drinking water in 1950s France.)





Practical 5c

PCA on Correlation Matrix

PCA on Correlation Matrix

To "scale" the data before PCA:

```
from sklearn.preprocessing import scale
Z = scale(X)
```

Example

(b) PCA on correlation matrix

p=7

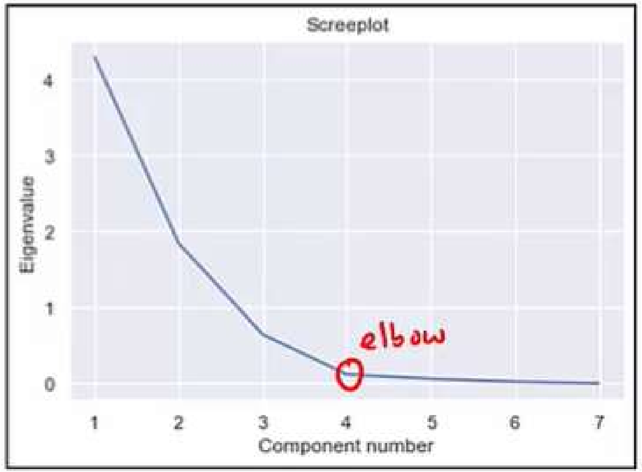
	Eigenvalue	Explained Variance	Cumulative Explained Variance	Bread	Vegetables	Fruit	Meat	Poultry	Milk	Wine
PC 1	4.2992	0.6142	0.6142	0.2324	0.4657	0.4505	0.4658	0.4355	0.2781	-0.2054
PC 2	1.8490	0.2641	0.8783	0.6259	0.0993	-0.1963	-0.1325	-0.1994	0.5193	0.4826
PC 3	0.6469	0.0924	0.9707	0.0181	-0.0829	0.1351	0.1979	0.3811	-0.4631	0.7587
PC 4	0.1205	0.0172	0.9879	-0.5633	0.0738	0.5375	-0.0999	-0.3144	0.3954	0.3510
PC 5	0.0613	0.0088	0.9967	-0.0214	0.8403	-0.0745	-0.3247	-0.1949	-0.3755	0.0590
PC 6	0.0218	0.0031	0.9998	0.4859	-0.2263	0.6551	-0.2071	-0.3243	-0.3392	-0.1430
PC 7	0.0013	0.0002	1.0000	-0.0112	-0.0610	0.1293	-0.7537	0.6191	0.1616	-0.0452

Total 7

Number of PCs to extract:

- By Kaiser's rule, extract the first 2 PCs whose eigenvalues (4.30 and 1.85) are > 1.
- 1st 2 PCs already accounted for 87.8% of total variance.
- Scree plot shows elbow at PC4; suggesting 1st 3 PCs to extract.

Let's extract the first 2 PCs only.



PCA on Correlation Matrix

Example

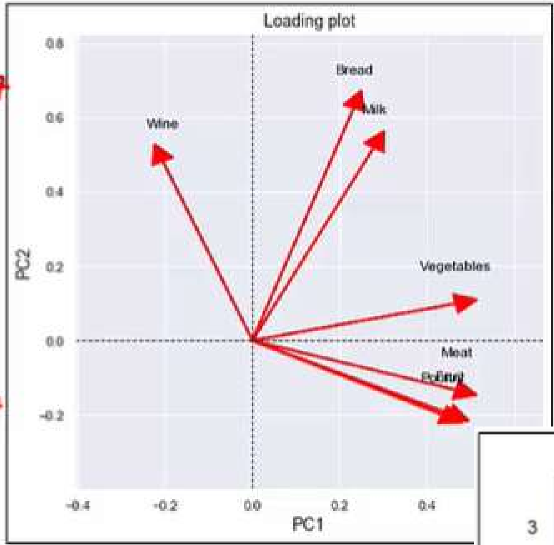
(b) PCA on correlation matrix

z_1 z_2 z_3 z_4 z_5 z_6 z_7

	Eigenvalue	Explained Variance	Cumulative Explained Variance	Bread	Vegetables	Fruit	Meat	Poultry	Milk	Wine
PC 1	4.2992	0.6142	0.6142	0.2324	0.4657	0.4505	0.4658	0.4355	0.2781	-0.2054
PC 2	1.8490	0.2641	0.8783	0.6259	0.0993	-0.1963	-0.1325	-0.1994	0.5193	0.4826

Bread Vegetables Fruit Meat Poultry Milk Wine

PC1: $\hat{y}_1 = 0.232 z_1 + 0.466 z_2 + 0.451 z_3 + 0.466 z_4 + 0.436 z_5 + 0.278 z_6 - 0.205 z_7$

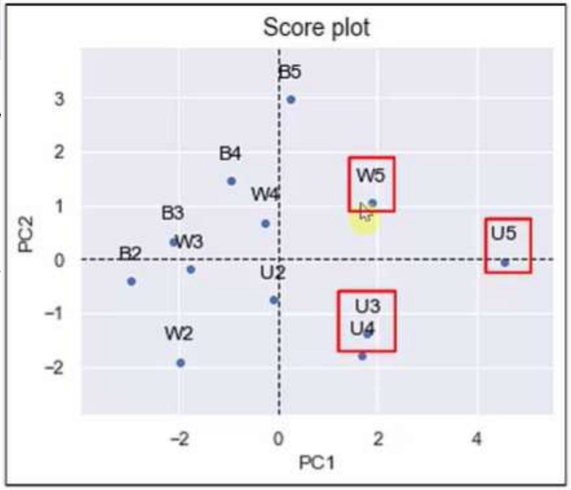


The loading on wine is opposite in sign to the other loadings.
This PC seems to measure a weighted average of expenditure on food, contrasting against wine.

Bread Vegetables Fruit Meat Poultry Milk Wine

PC2: $\hat{y}_2 = 0.626 z_1 + 0.099 z_2 - 0.196 z_3 - 0.133 z_4 - 0.199 z_5 + 0.519 z_6 + 0.483 z_7$

The loading on vegetables is quite small compared to other loadings.
The loadings on bread, milk and wine are opposite in sign to the loadings on fruit, meat and poultry.
This PC seems to measure a contrast of expenditure on “luxury” items against “non-luxury” items.



Example

PCA on covariance matrix

PC: $\hat{y} = 0.069x_1 + 0.327x_2 + 0.304x_3 + 0.756x_4 + 0.462x_5 + 0.090x_6 - 0.059x_7$

PCA on correlation matrix

PC1: $\hat{y}_1 = 0.232z_1 + 0.466z_2 + 0.451z_3 + 0.466z_4 + 0.436z_5 + 0.278z_6 - 0.205z_7$

PC2: $\hat{y}_2 = 0.626z_1 + 0.099z_2 - 0.196z_3 - 0.133z_4 - 0.199z_5 + 0.519z_6 + 0.483z_7$

Carry out PCA on correlation matrix when

- Variables have different units of measurement, or
- Variables have different scale or magnitude

PCA on covariance matrix can explain a higher percentage of total variance than PCA on correlation matrix, if same number of PCs are extracted.

Practical 5d

Example 2

Example 2.

Correlation matrix can be downloaded from Blackboard

In a psychological experiment, the reaction times of 64 normal men and women to visual stimuli were recorded when warning intervals of 0.5, 1, 3, 6, and 15 seconds preceded the stimulus. The correlations of the median reactions times of several replications of each preparatory interval for a subject formed this matrix:

$$\begin{pmatrix} 1 & 0.71 & 0.58 & 0.56 & 0.65 \\ 0.71 & 1 & 0.71 & 0.60 & 0.69 \\ 0.58 & 0.71 & 1 & 0.75 & 0.71 \\ 0.56 & 0.60 & 0.75 & 1 & 0.74 \\ 0.65 & 0.69 & 0.71 & 0.74 & 1 \end{pmatrix}$$

$n = 64, p = 5$

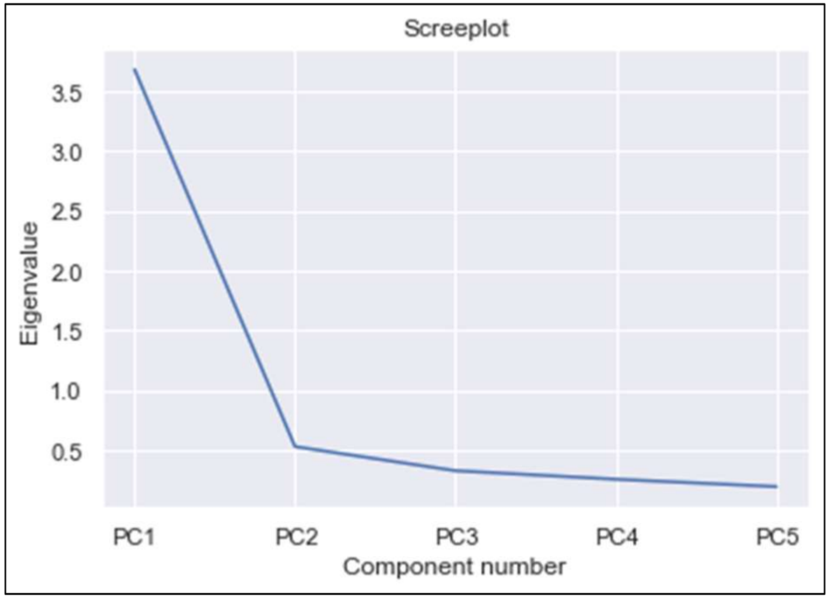
Extract the principal components and interpret the results.

	Eigenvalues	Explained Variance	Cumulative Explained Variance	0.5s	1s	3s	6s	15s
PC1	3.6831	0.7366	0.7366	0.4224	0.4506	0.4567	0.4439	0.4615
PC2	0.5313	0.1063	0.8429	0.6661	0.3784	-0.3255	-0.5359	-0.1414
PC3	0.3300	0.0660	0.9089	0.4030	-0.5657	-0.5151	0.3100	0.3951
PC4	0.2588	0.0518	0.9607	-0.4246	0.2817	-0.2883	-0.3470	0.7327
PC5	0.1969	0.0394	1.0001	0.1879	-0.5044	0.5805	-0.5470	0.2722

Number of PCs to extract:

- By Kaiser's rule, extract the 1st PC whose eigenvalue (3.68) is > 1.
- 1st 2 PCs already accounted for 84.3% of total variance. However,
  2nd PC account for 10.6% of total variance, which could be too high to discard.
- Scree plot shows elbow at PC2; suggesting 1 PC to extract.

Let's extract the first 2 PCs only.



	Eigenvalues	Explained Variance	Cumulative Explained Variance	0.5s	1s	3s	6s	15s
PC1	3.6831	0.7366	0.7366	0.4224	0.4506	0.4567	0.4439	0.4615
PC2	0.5313	0.1063	0.8429	0.6661	0.3784	-0.3255	-0.5359	-0.1414
PC3	0.3300	0.0660	0.9089	0.4030	-0.5657	-0.5151	0.3100	0.3951
PC4	0.2588	0.0518	0.9607	-0.4246	0.2817	-0.2883	-0.3470	0.7327
PC5	0.1969	0.0394	1.0001	0.1879	-0.5044	0.5805	-0.5470	0.2722

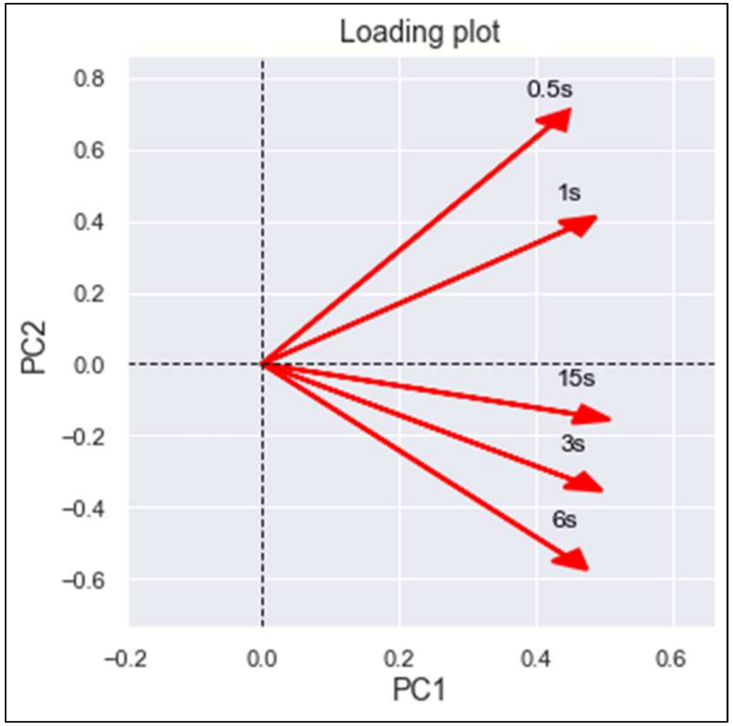
PC1: $\hat{y}_1 = 0.422z_1 + 0.451z_2 + 0.457z_3 + 0.444z_4 + 0.462z_5$

All the loadings are in the same direction. This PC seems to measure general reaction time.

PC2: $\hat{y}_2 = 0.666z_1 + 0.378z_2 - 0.326z_3 - 0.536z_4 - 0.141z_5$

The loadings on 0.5s and 1s are opposite in sign to the other loadings.

This PC seems to measure a contrast of reaction times preceding short stimuli, to reaction times preceding mid-to-long stimuli.



Practise Tutorial 5.2 before attempting Assignment 1