Chapter 5: Principal Component Analysis

Learning Objectives:

1. *State the purpose of principal component analysis.*
2. *Describe the steps involved in principal component analysis.*
3. *Apply the criteria for extracting the number of principal components.*
4. *Carry out principal component analysis on data sets and interpret principal components obtained.*

5.1 Introducing PCA

Suppose we have a multivariate data set with 10, 20 or even more variables. How are we going to visualize this data set when we can render plots only in 2-dimension, or at most 3-dimension? Is there, perhaps, a small number of *features* in this data set that are more *interesting* and can capture most of the information contained in the data set? The tool to achieve the above is ***Principal Component Analysis***.

5.1.1 What is PCA?

Principal component analysis (**PCA**) is an unsupervised learning approach, concerned with explaining the variance-covariance structure of a set of variables through a few linear combinations of these variables. These linear combinations of variables are called **principal components**. Hence, PCA is simply the process of deriving these principal components.

Since PCA is unsupervised, it is typically used as part of an exploratory data analysis. An extension of PCA is factor analysis. PCA can also serve as intermediate steps to supervised learning approaches. For example, principal components may be inputs to multiple regression analysis or linear discriminant analysis.

The general objectives of PCA are:

1. dimension reduction without much loss of information, and
2. data visualization and interpretation.

What does it mean to “reduce dimension without much loss of information”?

Let’s use a simple scenario to illustrate.
Suppose we want to investigate the price of necessary goods like bread and rice, so we sampled 100 shops and recorded the prices of two brands of bread and three brands of rice:

Shop	Bread Brand P	Bread Brand Q	Rice Brand A	Rice Brand B	Rice Brand C
#1	$2.15	$1.80	$10.25	$8.90	$11.95
#2	$2.10	$1.70	$10.00	$8.90	$12.05
#3	$2.10	$1.75	$10.10	$8.90	$12.00
…	…	…	…	…	…
#100	$2.20	$1.75	$10.50	$9.00	$12.50

Let's transform the data to this:

Shop	Average of Bread	Average of Rice
#1	\$1.98	\$10.37
#2	\$1.90	\$10.32
#3	\$1.93	\$10.33
…	…	…
#100	\$1.98	\$10.67

Originally, we have $n = 100$ observations and $p = 5$ variables.

After transforming:

- The new variables are linear combinations of the original variables. "Average of bread" is a linear combination of the prices of bread, whereas "average of rice" is a linear combination of the prices of rice.
- We now have $n = 100$ observations and $p = 2$ variables. So, we have reduced the dimension from $p = 5$ to $p = 2$.
- However, the information in the original data set cannot be re-created from the information in the transformed data set. Thus, information is lost after transformation.

5.1.2 What are Principal Components?

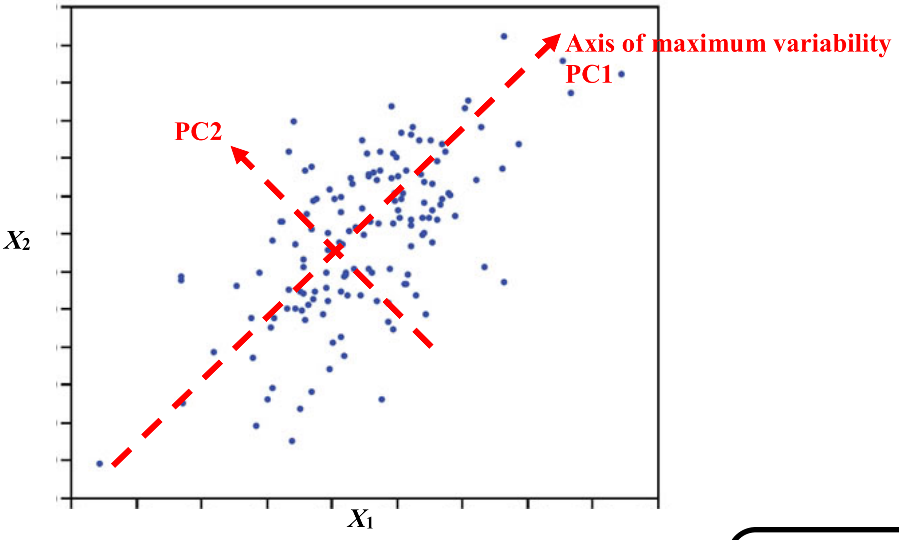
Principal components are uncorrelated linear combinations (of variables) whose variances are as large as possible. So in a data set with many correlated variables, a small number of principal components will be sufficient to explain most of the variability contained in this data set.

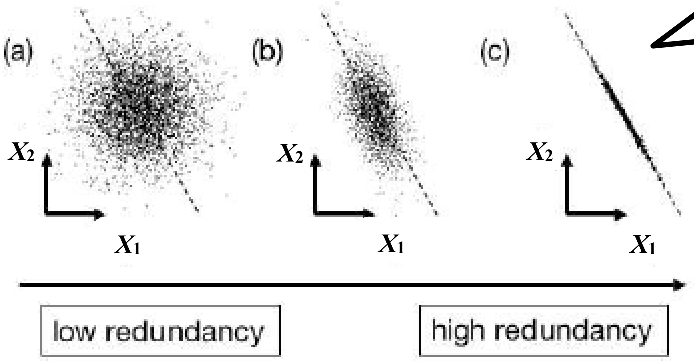
To elaborate further, although p variables or components are required to reproduce the total system variability, often much of this variability can be accounted for by a small number m of the principal components ($m < p$). If so, there is almost as much information in the m principal components as there is in the original p variables. The m principal components can then replace the initial p variables, and the original data set consisting of n measurements on p variables, is reduced to a data set consisting of n measurements on m principal components.

How do the principal components retain most of the information in terms of variability?

The figure on the next page depicts a bivariate data point of view:

- The first principal component (PC1) will be in the direction of the axis of maximum variability in the data set.
- The second principal component (PC2) will be in the direction perpendicular to the axis of maximum variability so as to capture variability not captured by PC1.
- PCA will transform the data set by rotation such that the axis of maximum variability becomes the new horizontal axis, and the axis perpendicular to the axis of maximum variability becomes the new vertical axis. This rotation is called **varimax rotation**.





Which of these data sets are suitable for dimension reduction?

5.2 Assumptions and Limitations

We list here some important assumptions and limitations of PCA.

- Sample Size:
  Correlation coefficients tend to be less reliable when estimated from small samples. Therefore, it is important that sample size be large enough that correlations are reliably estimated.

- Normality:
  PCA is a generally non-parametric analysis. If variables are normally distributed, the solution is enhanced. To the extent that normality fails, the solution is degraded but may still be worthwhile.

- Linearity:
  The analysis is degraded when linearity fails, because correlation measures linear relationship and does not reflect non-linear relationship.

- Orthogonality:
  Principal components are orthogonal.

5.3 Illustrating with Bivariate Data

We will use the following bivariate data to guide us through the process of PCA:

X_1	2.5	0.5	2.2	1.9	3.1	2.3	2.0	1.0	1.5	1.1
X_2	2.4	0.7	2.9	2.2	3.0	2.7	1.6	1.1	1.6	0.9

(*Data for this example can be downloaded from Blackboard.*)

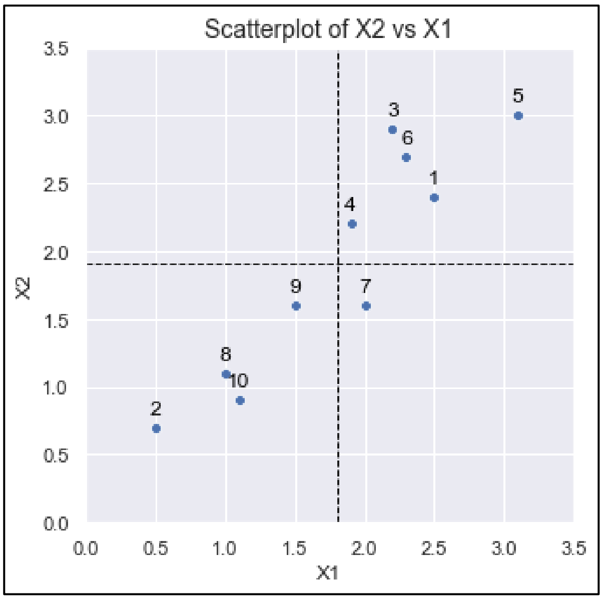
Construct a scatterplot to see how the data is distributed →

Centre: $\overline{\mathbf{x}} = \begin{pmatrix} \quad \\ \quad \end{pmatrix}$

Notice that X_1 and X_2 are positively correlated.

$r =$

(Try sketching an ellipse around the data points.)

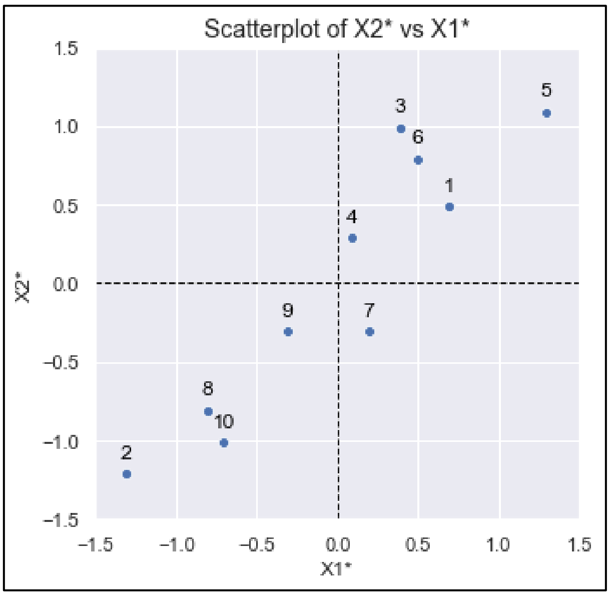


STEP 1: Subtract the means from the corresponding data component to re-centre the data set. Re-construct the scatterplot to view.
Write the "adjusted" data as a matrix $\mathbf{X}^*$. Note that the "adjusted" data set will have means zero.

X_1^*										
X_2^*										

$\mathbf{X}^* =$

$\overline{\mathbf{x}}^* =$



<u>STEP 2</u>: Compute the sample covariance matrix **C**.

$$\mathbf{C}=\frac{1}{n-1}\left(\mathbf{X}-\mathbf{1}_n\overline{\mathbf{x}}'\right)'\left(\mathbf{X}-\mathbf{1}_n\overline{\mathbf{x}}'\right)=\frac{1}{n-1}\mathbf{X}^{*}{}'\,\mathbf{X}^{*}$$

If the data is standardized, then **C** is the correlation matrix: $\mathbf{C}=\frac{1}{n-1}\mathbf{Z}'\mathbf{Z}$

STEP 3: Compute the eigenvalues λ_i, and normalized eigenvectors $\hat{\mathbf{v}}_i$ of **C**, order the corresponding pairs from the highest to the lowest eigenvalues.

Using the property that the sum of the eigenvalues of a matrix equals its trace,

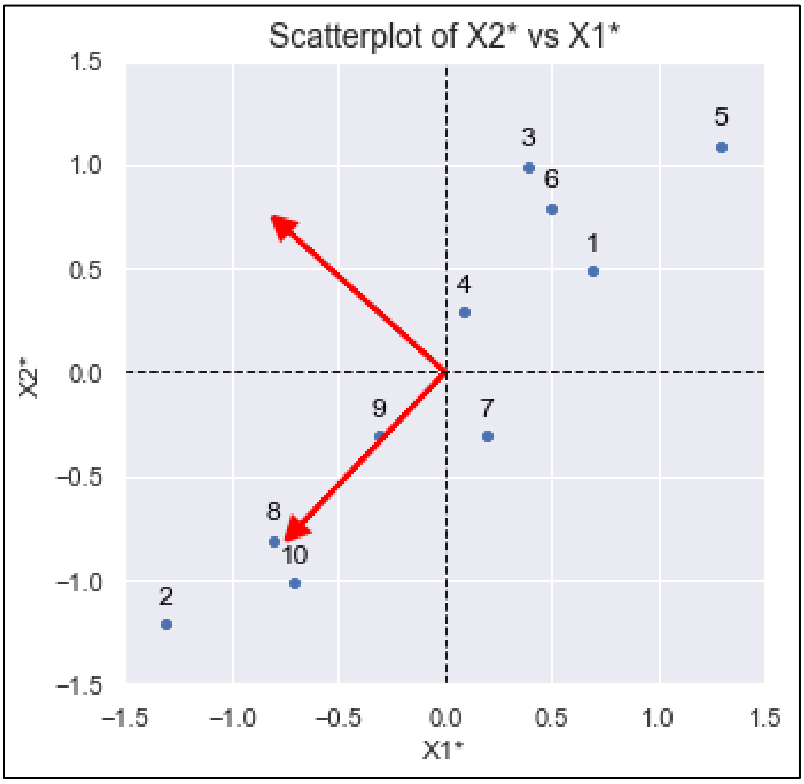
total sample variance = tr(C) = sum of eigenvalues of C

By this process, we will be able to extract axes that characterize the data.

The first eigenvector will go through the middle of the data points, as if it is the line of best fit. This line is the axis of maximum variability, giving the direction of PC1.

The second eigenvector will give us the other, less important, pattern in the data. That is, all the data points follow the main line, but are off to the side of the main line by some amount. This will be along the axis perpendicular to PC1, capturing variability not captured by PC1, thus giving the direction of PC2.

Variable	Eigenvector 1	Eigenvector 2
X_1^* X_2^*	–0.678 –0.735	–0.735 0.678
Eigenvalues	1.284	0.049
% of total variance		



STEP 4: Choose the components and form the eigenvector matrix **V**.

By ordering the eigenvectors according to the eigenvalues, this gives the components in order of their significance. Hence, the eigenvector with the highest eigenvalue is the principal component. The components of lesser significance can be ignored, so as to reduce the dimensions of the data set.

Then in this example, we can either, …

- Select both components, then **V** =

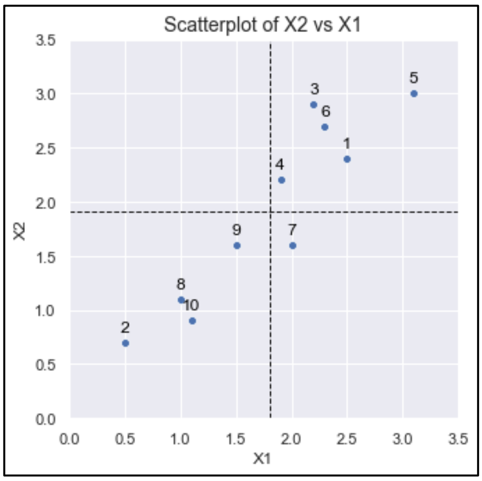
- Or, discard the less significant component, then **V** =

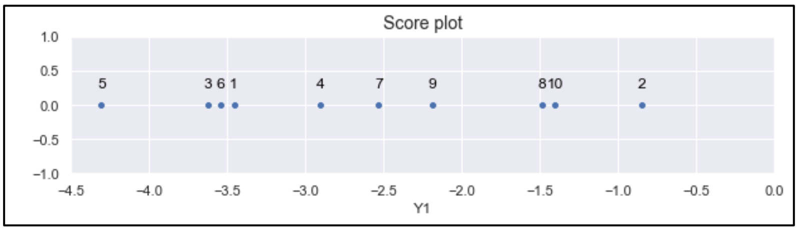
STEP 5: Derive the new data set by taking **Y** = **XV**.

Basically we have transformed our data so that it is expressed in terms of the patterns between them, where the patterns are the lines or axes that most closely describe the relationships between the data.

Supposing that we only retain PC1 and discard PC2. The transformed data **Y**, known as **scores**, will then be univariate.

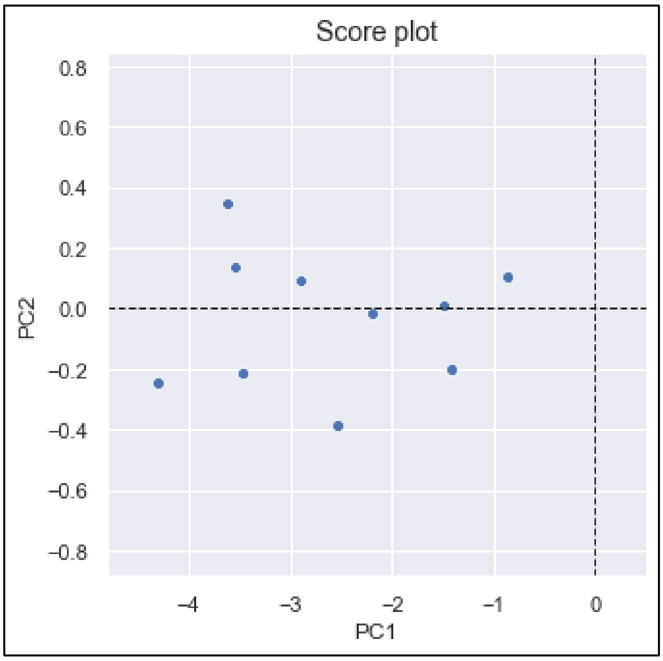
The data points in **X** are plotted in the left figure below, and the scores are plotted in the right figure below.



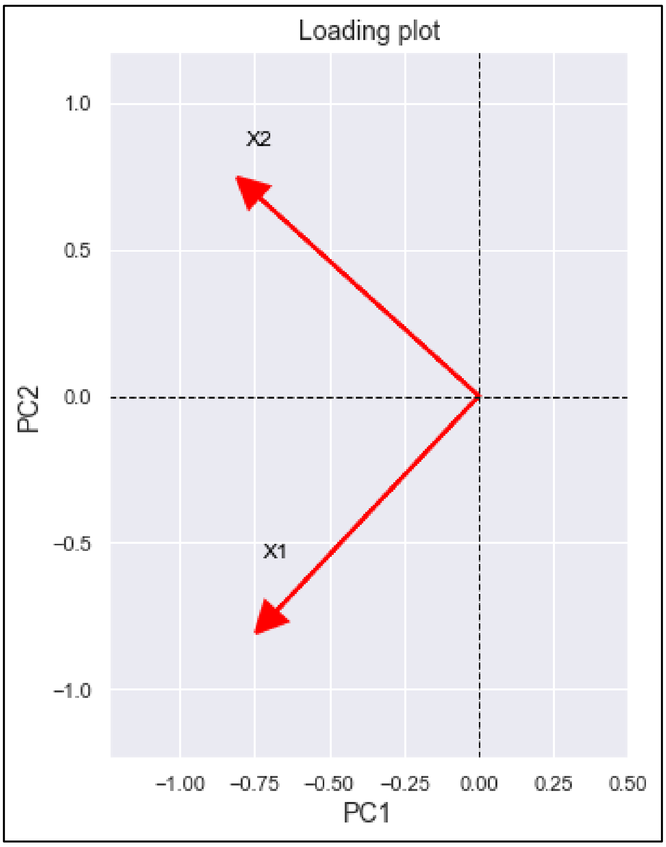


Practical 5a

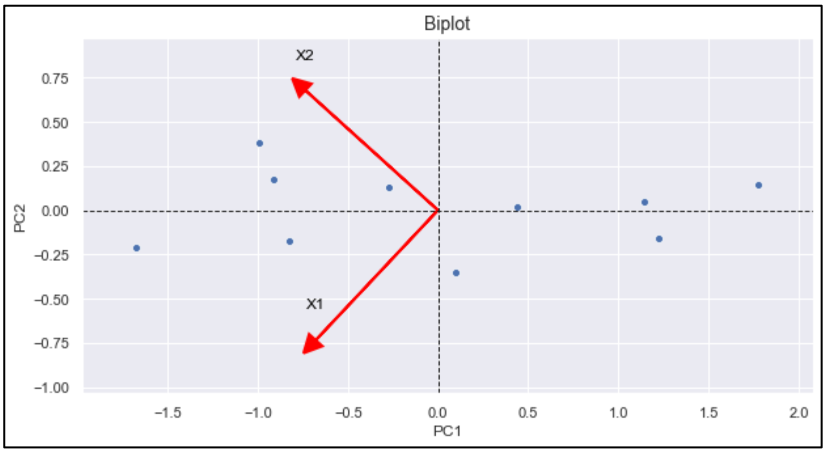
The following charts are commonly used in PCA. We will examine how to plot them using Python.



This **score plot** shows



This **loading plot** shows



This **biplot** is

5.4 Extracting Principal Components

In multivariate data with *p*-variables, there will be *p* principal components. We will want to retain only the first few PCs which can capture as much variability as possible. How many PCs is sufficient? Unfortunately, there is no objective method to determine how many PCs is needed. It depends on the specific area of application and specific data set.

Nevertheless, here are three criteria that can help you decide on the number of PCs to extract. Note that these are just guidelines, not theorems.

#1. Look at the eigenvalues

If PCA is performed on the correlation matrix, the *p* eigenvalues should sum up to *p*. Kaiser felt that a PC whose eigenvalue is less than 1 explains less variance that a variable would, and should be discarded. Thus, Kaiser's rule state that we should retain PCs whose eigenvalues are greater than 1. Although this is a simple rule, the decision is not simple when eigenvalues are close to 1, for example, 1.002 and 0.998.

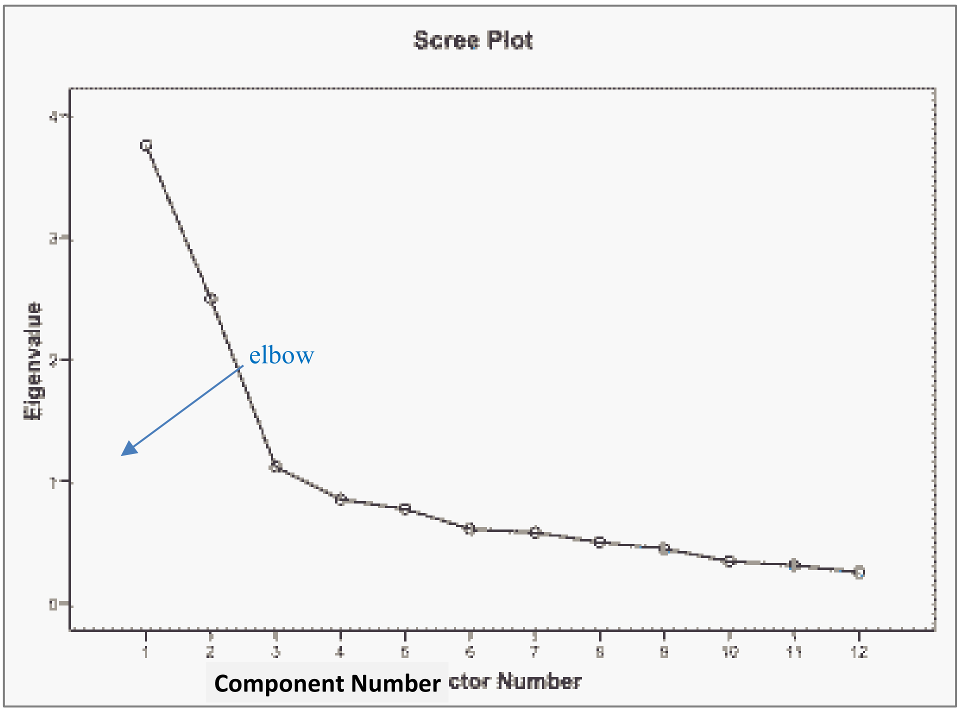
#2. Look at the percentage of total variance

The first PC is in the direction of maximum variability, thus it should explain the highest percentage of total variation. This will be followed by PC2, PC3 and so on. It is suggested that PCs which explain less than 5−10% of the total variability should be discarded. Another suggestion is to retain PCs with cumulative percentage of at least 70−80% if intention of PCA is descriptive; retain PCs with cumulative percentage of at least 80−90% if PCA is intended to lead to further analysis. As you can see, the weakness of this criteria is that the percentage figures to determine number of PCs to extract can be arbitrary and subjective.

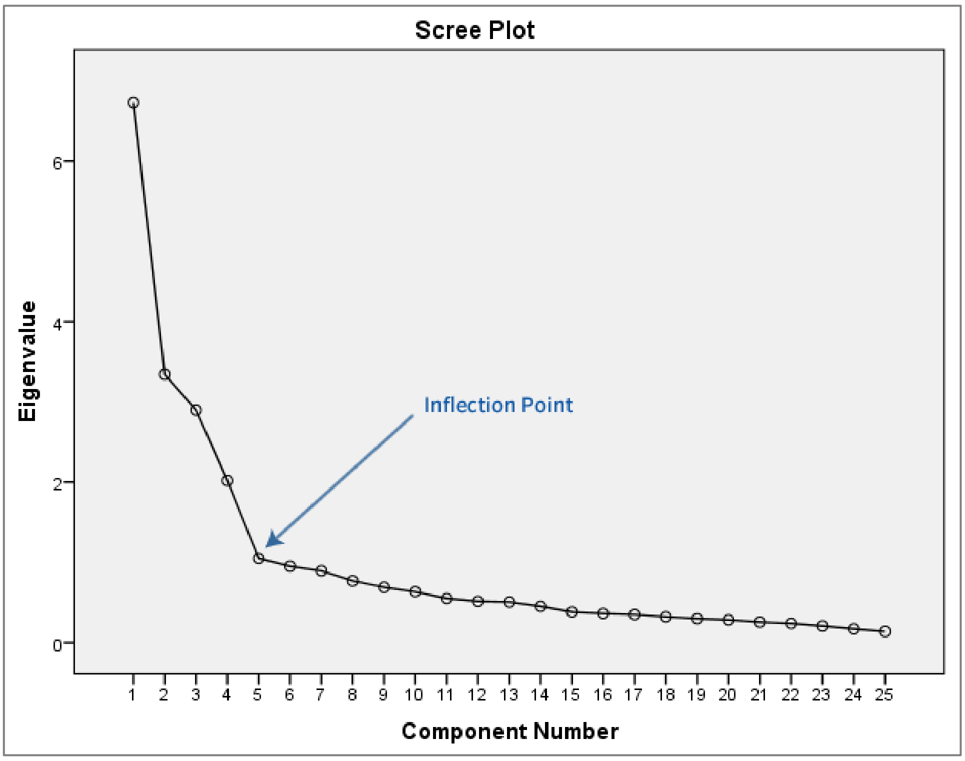
#3. Look at a scree plot

A scree plot is a line graph with eigenvalues on the vertical axis and PC number on the horizontal axis, arranged in descending eigenvalues. Usually, an inflection point will show up in the graph; this is the point where the graph begins to "flatten" and subsequent components add little to the total variance. This inflection point is called the **elbow** in the plot. It is suggested to retain PCs on the left of the elbow. A couple of samples of scree plots can be found in the next page.

You will notice that the number of PCs suggested to retain using the different criteria are not always the same. Hence, the decision is subjective.



This scree plot suggests retaining 2 PCs.



This scree plot suggests retaining 4 PCs.

5.5 Multivariate Examples

Example 1.

Data can be downloaded from Blackboard

The context of the data collected below is set in the 1950s in France. The average number of Francs spent on several categories of food products according to social class and the number of children per family was recorded.

Social Class/ No. of children	Bread	Vegetables	Fruit	Meat	Poultry	Milk	Wine
B2	332	428	354	1437	526	247	427
W2	293	559	388	1527	567	239	258
U2	372	767	562	1948	927	235	433
B3	406	563	341	1507	544	324	407
W3	386	608	396	1501	558	319	363
U3	438	843	689	2345	1148	243	341
B4	534	660	367	1620	638	414	407
W4	460	699	484	1856	762	400	416
U4	385	789	621	2366	1149	304	282
B5	655	776	423	1848	759	495	486
W5	548	995	548	2056	839	518	319
U5	515	1097	887	2630	1167	561	284

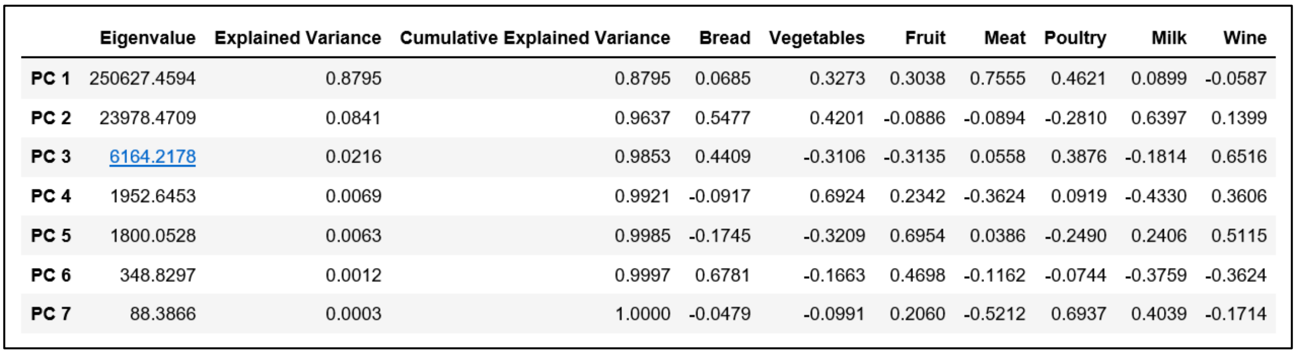
Legend
B: blue-collared
W: white-collared
U: upper-class

Extract the principal components and interpret the results using the:

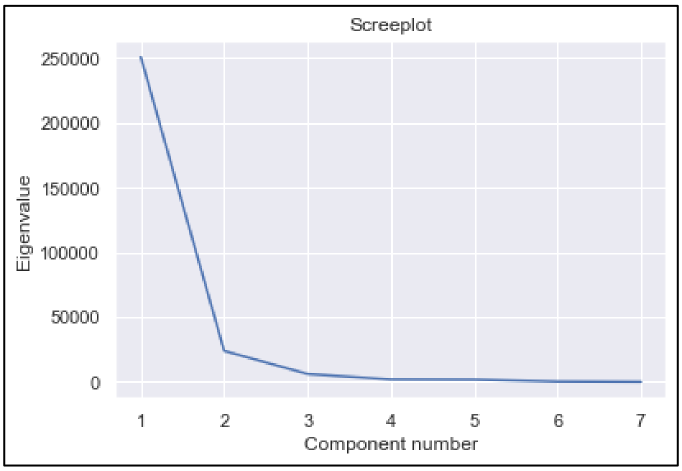
(a) covariance matrix, (b) correlation matrix.

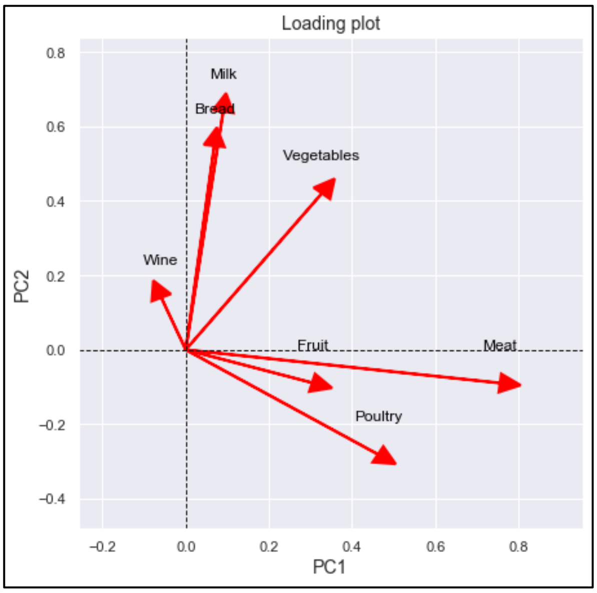
$n =$ _____ , $p =$ _____

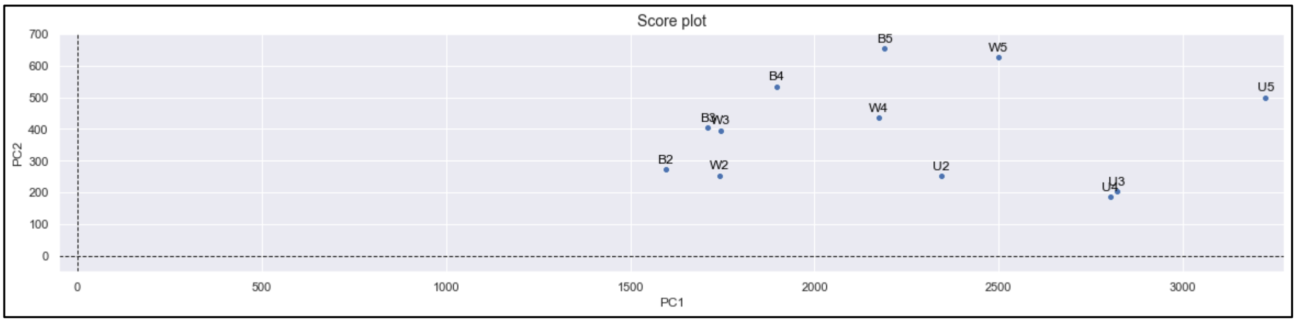
(a) PCA on covariance matrix:



	Eigenvalue	Explained Variance	Cumulative Explained Variance	Bread	Vegetables	Fruit	Meat	Poultry	Milk	Wine
PC 1	250627.4594	0.8795	0.8795	0.0685	0.3273	0.3038	0.7555	0.4621	0.0899	-0.0587
PC 2	23978.4709	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

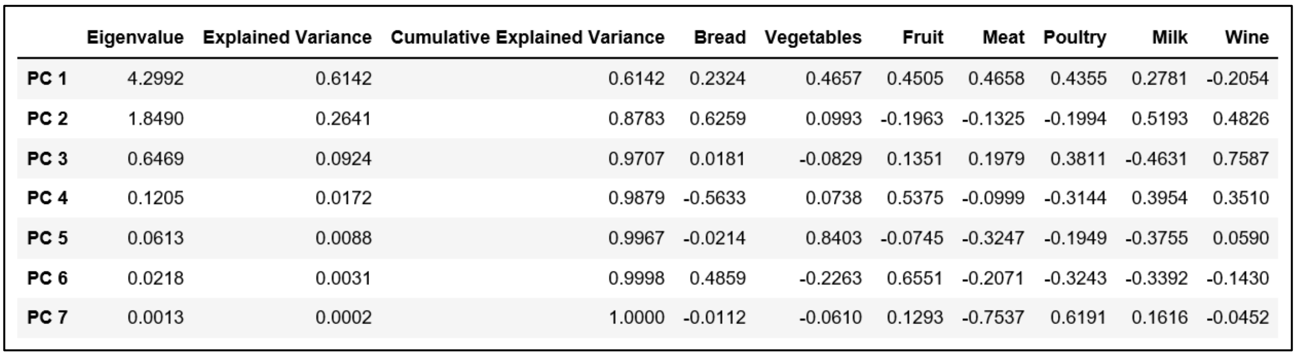




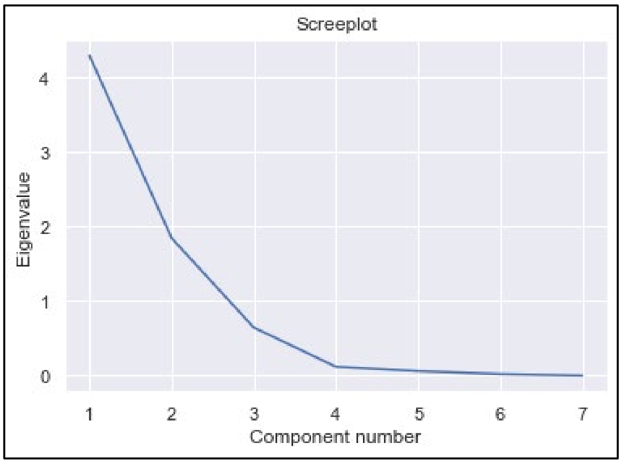


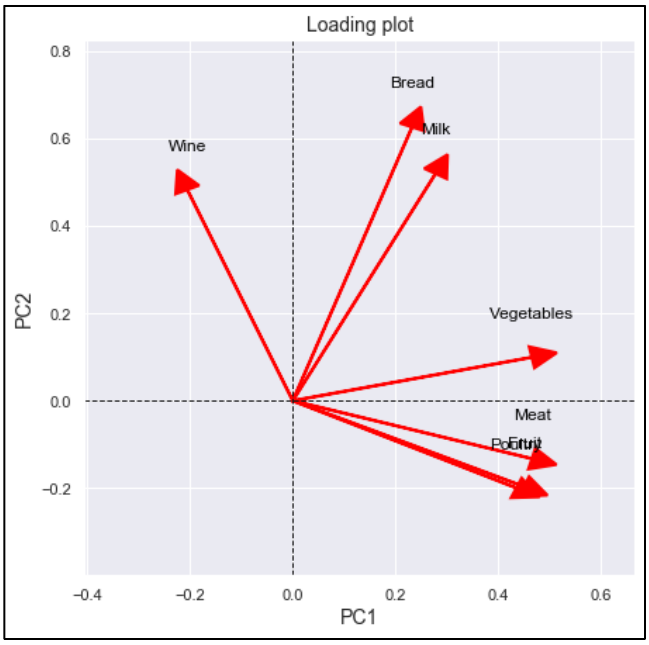
Practical 5b: Modify the functions used in Practical 5a to obtain the four outputs above.

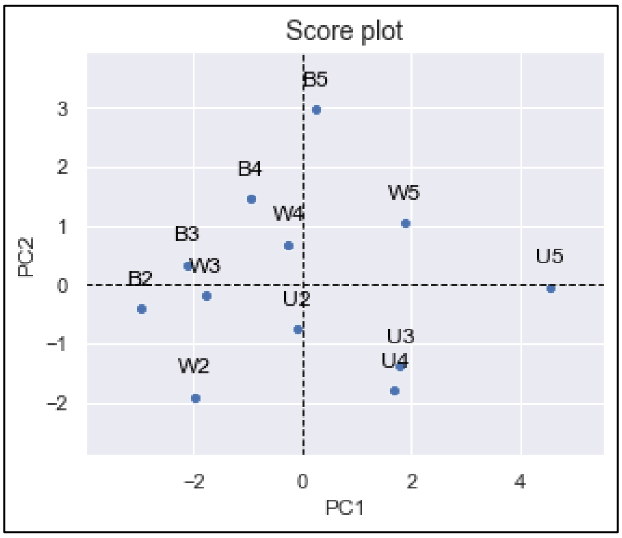
(b) PCA on correlation matrix:



	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







Practical 5c: Use the functions developed in Practical 5b to obtain the four outputs above.

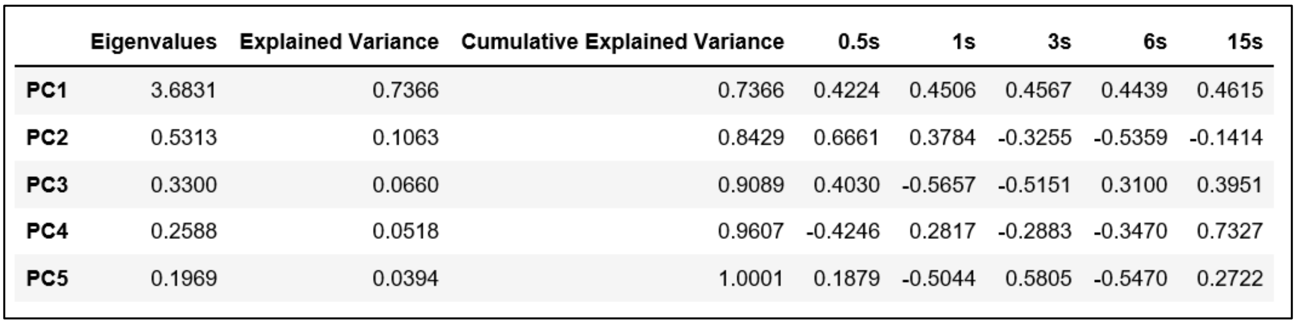
Example 2.

Correlation matrix can be downloaded from Blackboard

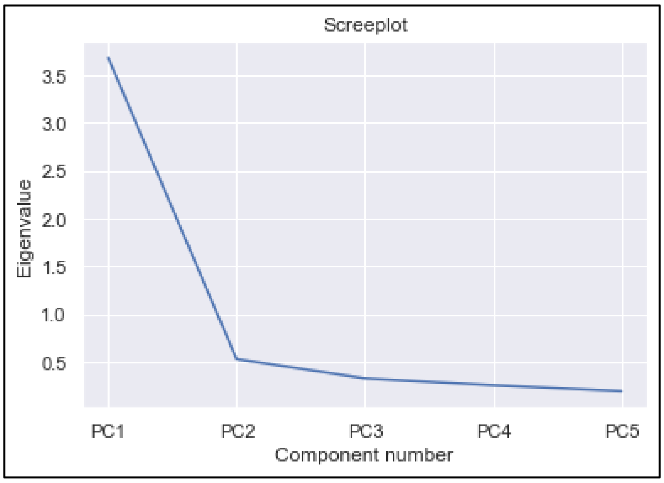
In a psychological experiment, the reaction times of 64 normal men and women to visual stimuli were recorded when warning intervals of 0.5, 1, 3, 6, and 15 seconds preceded the stimulus. The correlations of the median reactions times of several replications of each preparatory interval for a subject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

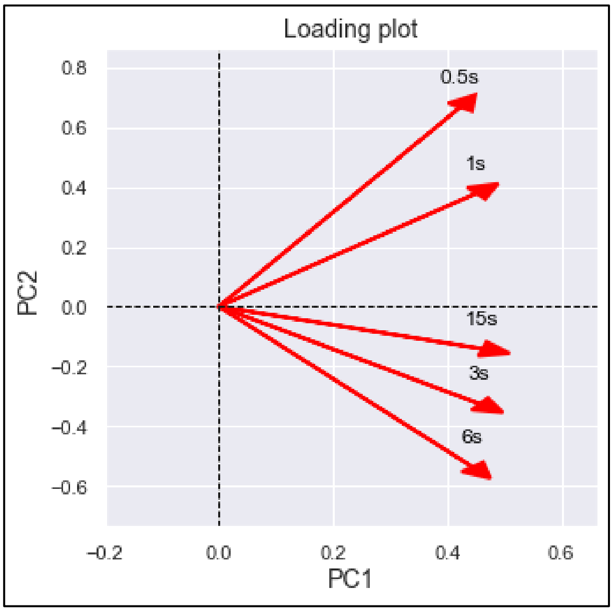
$$\begin{pmatrix} 1 & 0.71 & 0.58 & 0.56 & 0.65 \\ 0.71 & 1 & 0.71 & 0.60 & 0.69 \\ 0.58 & 0.71 & 1 & 0.75 & 0.71 \\ 0.56 & 0.60 & 0.75 & 1 & 0.74 \\ 0.65 & 0.69 & 0.71 & 0.74 & 1 \end{pmatrix}$$

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





Practical 5d: Use the eigenvalues and eigenvectors dataframes to obtain the three outputs above.

Tutorial 5.1

1. The sample covariance matrix for a bivariate data set is

$$\mathbf{C} = \begin{bmatrix} 5 & 2 \\ 2 & 2 \end{bmatrix}$$

Find the first principal component. What is the percentage of total variance explained by the first principal component?

2. Convert the sample covariance matrix in question 1 to sample correlation matrix **R**.

(a) Determine the first principal component. Compute the percentage of total variance explained by the first principal component.

(b) Compare the percentage of total variance explained by the first principal component in question 1 with that obtained in part (a).

3. Data on X_1 = sales and X_2 = profits for 10 large companies in the world are obtained as follow:

Company	X_1 **($ billion)**	X_2 **($ billion)**
Citgroup	108.28	17.05
General Electric	152.36	16.59
AIG	95.04	10.91
Bank of America	65.45	14.14
HSBC Group	62.97	9.52
ExxonMobil	263.99	25.33
Royal Dutch/Shell	265.19	18.54
BP	285.06	15.73
ING Group	92.01	8.1
Toyota Motor	165.68	11.13

Data can be downloaded from Blackboard

They have the following sample mean and sample covariance matrix:

$$\overline{\mathbf{x}} = \begin{bmatrix} 155.60 \\ 14.70 \end{bmatrix} \qquad \mathbf{C} = \begin{bmatrix} 7476.45 & 303.62 \\ 303.62 & 26.19 \end{bmatrix}$$

(a) Using the sample covariance matrix **C**,

(i) determine the two principal components and their eigenvalues.

(ii) find the percentage of total variance explained by each principal component.

(iii) Create a scatterplot of the data. Indicate the directions of the principal components obtained on your scatterplot.

(b) Compute the sample correlation matrix **R** of the data. Then,

(i) determine the two principal components and their eigenvalues.

(ii) find the percentage of total variance explained by each principal component.

(iii) Create a scatterplot of the data. Indicate the directions of the principal components obtained on your scatterplot.

(c) Compare your results obtained in parts (a) and (b).

Tutorial 5.2

1. The Madison, Wisconsin, police department regularly monitors many of its activities as part of an on-going quality improvement program. The partial table below gives the data on five different kinds of overtime hours:

Legal Appearance Hours	Extraordinary Event Hours	Holdover Hours	Compensatory Overtime Allowed Hours	Meeting Hours
X_1	X_2	X_3	X_4	X_5
3387	2200	1181	14861	236
3109	875	3532	11367	310
2670	957	2502	13329	1182
…	…	…	…	…

Data can be downloaded from Blackboard

Each observation represents a total for 12 pay periods, or about half a year.
Extract the principal components and interpret the results.

2. *Another Biometric Application: Fowl-Bone Lengths*. Wright (1954) reported these correlations among six bone dimensions of $n = 276$ white Leghorn fowl:

$$\begin{array}{r} \text{Skull length} \\ \text{Skull breadth} \\ \text{Humerus} \\ \text{Ulna} \\ \text{Femur} \\ \text{Tibia} \end{array} \begin{pmatrix} 1 & 0.584 & 0.615 & 0.601 & 0.570 & 0.600 \\ 0.584 & 1 & 0.576 & 0.530 & 0.526 & 0.555 \\ 0.615 & 0.576 & 1 & 0.940 & 0.875 & 0.878 \\ 0.601 & 0.530 & 0.940 & 1 & 0.877 & 0.886 \\ 0.570 & 0.526 & 0.875 & 0.877 & 1 & 0.924 \\ 0.600 & 0.555 & 0.878 & 0.886 & 0.924 & 1 \end{pmatrix}$$

(*This correlation matrix can be downloaded from Blackboard.*)

The skull contributes two dimensions, while two wing bones (humerus and ulna) and two leg bones (femur and tibia) are also represented.

Perform principal component analysis on the correlation matrix and interpret the results.

3. Jolicoeur and Mosimann (1960) have investigated the principal components of carapace length, width, and height of painted turtles in an effort to give meanings to the concepts of "size" and "shape". The covariance matrix of the lengths, widths, and heights in millimetres of the carapaces of 24 female turtles was:

$$\mathbf{C} = \begin{pmatrix} 451.39 & 271.17 & 168.70 \\ 271.17 & 171.73 & 103.29 \\ 168.70 & 103.29 & 66.65 \end{pmatrix}$$

Perform principal component analysis on the covariance matrix and interpret the results.

4. To study the cost of food items, data on prices of five food items were collected from 23 cities in the United States. Table below shows part of the data collected.

	Average Price (in cents per pound)				
City	**Bread**	**Burger**	**Milk**	**Oranges**	**Tomatoes**
Atlanta	24.5	94.5	73.9	80.1	41.6
Baltimore	26.5	91.0	67.5	74.6	53.3
⋮	⋮	⋮	⋮	⋮	⋮
Washington, DC	24.2	93.8	66.0	81.6	46.2

(a) Should PCA be carried out on covariance or correlation matrix? Explain.

(b) How many principal components should be extracted? Justify your answer.

(c) Write down the extracted principal components and interpret them.

(d) Based on the five food items collected, name the most expensive city and the least expensive city in food prices.

Answers

Tutorial 5.1

1. $\hat{y}_1 = 0.894x_1 + 0.447x_2$, 85.7%

2. $\begin{bmatrix} 1 & 0.6325 \\ 0.6325 & 1 \end{bmatrix}$

   (a) $\hat{y}_1 = 0.707x_1 + 0.707x_2$, 81.6%

   (b) A higher percentage of total variance is explained by PC1 in question 1, compared to in question 2. Generally, we will obtain different PCs and percentages of total variance explained by each PC when PCA is carried out on sample covariance matrix versus on a sample correlation matrix.

3. (a) (i) $\hat{y}_1 = 0.999x_1 + 0.041x_2$, $\hat{\lambda}_1 = 7488.8$

   $\hat{y}_2 = -0.041x_1 + 0.999x_2$. $\hat{\lambda}_2 = 13.8$

   (ii) 99.8%, 0.2%

   (b) $\begin{bmatrix} 1 & 0.6861 \\ 0.6861 & 1 \end{bmatrix}$

   (i) $\hat{y}_1 = 0.707z_1 + 0.707z_2$, $\hat{\lambda}_1 = 1.686$

   $\hat{y}_2 = -0.707z_1 + 0.707z_2$, $\hat{\lambda}_2 = 0.314$

   (ii) 84.3%, 15.7%

   (c) The PCs and percentage of total variance explained by each PC is different. When PCA is carried out on covariance matrix, the first PC explains almost all the total variance.

Tutorial 5.2 (Short answers)

1. PCA on correlation matrix.
   Extract 3 PCs which account for 86.7% of the variability.
   PC1: $\hat{y}_1 = -0.156z_1 + 0.074z_2 + 0.599z_3 - 0.577z_4 + 0.528z_5$
   PC1 measures a contrast of Legal Appearance and Compensatory Overtime hours against Holdover and Meeting overtime hours.
   PC2: $\hat{y}_2 = -0.648z_1 + 0.677z_2 - 0.286z_3 + 0.114z_4 + 0.163z_5$
   PC2 measures a contrast of Legal Appearance and Holdover overtime hours against Extraordinary Event, Compensatory Overtime and Meeting overtime hours.
   PC3: $\hat{y}_3 = 0.661z_1 + 0.607z_2 + 0.200z_3 + 0.319z_4 + 0.233z_5$
   PC3 is a measure of the overall overtime hours.

2. Extract 2 PCs which account for 88.0% of the variability.
   PC1: $\hat{y}_1 = -0.347z_1 - 0.326z_2 - 0.443z_3 - 0.440z_4 - 0.435z_5 - 0.440z_6$
   PC1 measures the overall rating of the bone dimensions.
   PC2: $\hat{y}_2 = -0.537z_1 - 0.697z_2 + 0.187z_3 + 0.251z_4 + 0.278z_5 + 0.226z_6$
   PC2 measures the contrast of skull dimensions to the other dimensions.

3. Extract 1 PC which account for 98.6% of the variability.

   PC: $\hat{y} = -0.813x_1 - 0.496x_2 - 0.307x_3$

   PC1 is a measure of the overall size of carapace.

4. (a) PCA should be carried out on correlation matrix as average prices of the five food items have different magnitudes.

   (b) Extract the first three PCs which accounts for 85.3% of total variance.

   (c) PC1: $\hat{y}_1 = 0.496z_1 + 0.576z_2 + 0.340z_3 + 0.225z_4 + 0.506z_5$

   PC1 is a weighted average of all food prices.

   PC2: $\hat{y}_2 = 0.309z_1 + 0.044z_2 + 0.431z_3 - 0.797z_4 - 0.287z_5$

   PC2 measures a contrast of prices of fruit / vegetable items vs non-fruit / non-vegetable items.

   PC3: $\hat{y}_3 = -0.386z_1 - 0.263z_2 + 0.835z_3 + 0.292z_4 - 0.012z_5$

   PC3 measures a contrast of prices of staple vs non-staple foods of people in the United States.

   (d) Most expensive: Honolulu

   Least expensive: Seattle