DM end term exam 1 Problem 1

Import libraries

In [3]:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
```

Parameters

In [4]:

```
csv_in = 'dm-end1-1.csv'
```

Read CSV file

In [5]:

```
df = pd.read_csv(csv_in, delimiter=',', skiprows=0, header=0)
print(df.shape)
print(df.info())
display(df.head())
```

```
(40, 5)
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 40 entries, 0 to 39
Data columns (total 5 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Jpn     40 non-null     int64
 1   Eng     40 non-null     int64
 2   Math    40 non-null     int64
 3   Phys    40 non-null     int64
 4   Chem    40 non-null     int64
dtypes: int64(5)
memory usage: 1.7 KB
None
```

	Jpn	Eng	Math	Phys	Chem
0	59	57	50	54	46
1	43	45	47	50	48
2	48	42	57	57	57
3	46	46	60	61	54
4	40	36	31	32	36

Set data

In [6]:

```
dfX = df
print(dfX.shape)
display(dfX.head())
```

```
(40, 5)
```

	Jpn	Eng	Math	Phys	Chem
0	59	57	50	54	46
1	43	45	47	50	48
2	48	42	57	57	57
3	46	46	60	61	54
4	40	36	31	32	36

Standardization

In [7]:

```
sc = StandardScaler()
X_std = sc.fit_transform(dfX)
```

PCA

In [8]:

```
n_pca = 5
pca = PCA(n_components=n_pca)
X_pca = pca.fit_transform(X_std)
```

Check contribution ratio

In [9]:

```
print(pca.explained_variance_ratio_)
print(np.cumsum(pca.explained_variance_ratio_))
```

```
[0.64671026 0.21353731 0.06909157 0.03560072 0.03506014]
[0.64671026 0.86024757 0.92933913 0.96493986 1.        ]
```

Ans. 1,

0.647
0.214

2

2D plot

Draw biplot

In [10]:

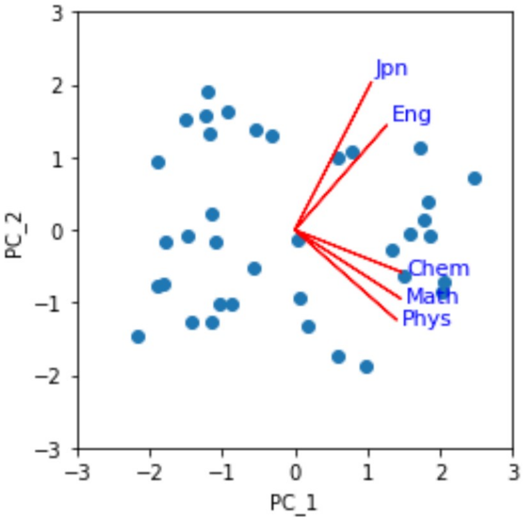
```
def biplot(X_2d, coef_2d, coef_labels=None):
    r1 = 3.
    r2 = 1.05
    coef_2dT = coef_2d.T
    if coef_labels is None:
        coef_labels = range(len(coef_2dT))
    for i, coef in enumerate(coef_2dT):
        plt.arrow(0, 0, coef[0]*r1, coef[1]*r1, color='r')
        plt.text(coef[0]*r1*r2, coef[1]*r1*r2, coef_labels[i],
                 color='b', fontsize=11)
    plt.scatter(X_2d[:,0], X_2d[:,1])
    plt.xlabel('PC_1')
    plt.ylabel('PC_2')

    plt.xlim(-3,3)
    plt.ylim(-3,3)
    plt.gca().set_aspect('equal', adjustable='box')

    return None
```

In [11]:

```
biplot(X_pca[:, :2], pca.components_[:2], coef_labels=dfX.columns)
```



第2主成分軸は、国語や英語といった言語科目の成績と正に相関している
The second principal component axis is positively correlated with the performance of language subjects such as Japanese and English.

In [ ]: