Python seminar Homework for Chap. 7.3 No.6

Please conduct (PCA) principal component analyses (empirical orthogonal function analyses: EOF analyses) on wet weight time series of fishes around Japan.

We will conduct two types of analyses: one is using raw data and the other is using standardized data (average is 0 and standard deviation is 1).

The raw wet weight time series of 6 fish stocks are in “weight.csv”.
The data is from 1978 to 2018.
The header has xxxY_Z format, where xxx is fish species, Y is regional stock and Z is life stage.
For species,

maiwashi: Japanese sardine
masaba: chub mackerel
urumeiwashi: round herring
katakuchi: Japanese anchovy

For regional stock,

P: Pacific
T: Tsushima

For life stage,

J: juvenile
m: juvenile to mature mixed
M: mature

For principal component analysis, there are two types: using covariance or correlation matrix.
For objectives to find out leading signals including absolute variational amplitude, covariance is useful, but the unit of data should be same (cannot apply for temperature and salinity combined data).
For objectives to find out pattern of variation, correlation is useful and this method can be applied to different unit mixed data.

scikit-learn (sklearn) is a tool to conduct machine learning. To conduct PCA, please import sklearn by

```
import sklearn
```

1. Please conduct PCA based on covariation matrix (using raw wet weight data).

PCA can be applied by the following commands.

```
pca = PCA()
pca.fit(data)
```

Read out explained variance ratio of PCs from pca by the following command.

```
pd.DataFrame(pca.explained_variance_ratio_, index=["PC{}".format(x + 1) for x in range(len(data.columns))])
```

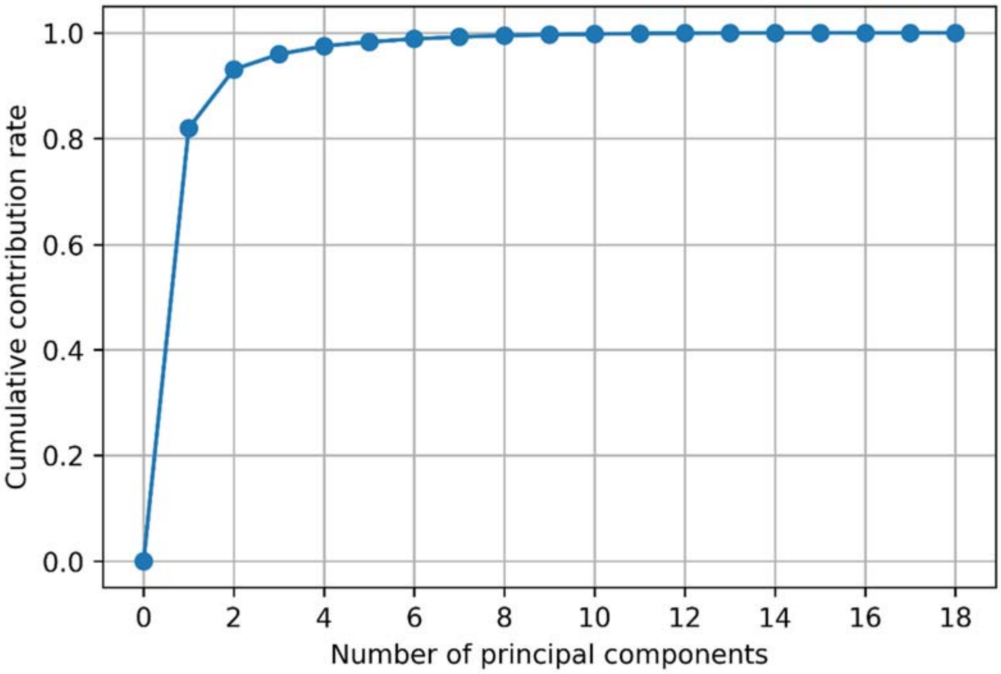

Figure 1. Cumulative contribution rate of PCs. In this case, PC1 explained more than 80% of the total variance.

Read out eigen vector of each PC from pca by the following command.

```
pd.DataFrame(pca.components_, columns=data.columns[:], index=["PC{}".format(x + 1) for x in range(len(data.columns))])
```

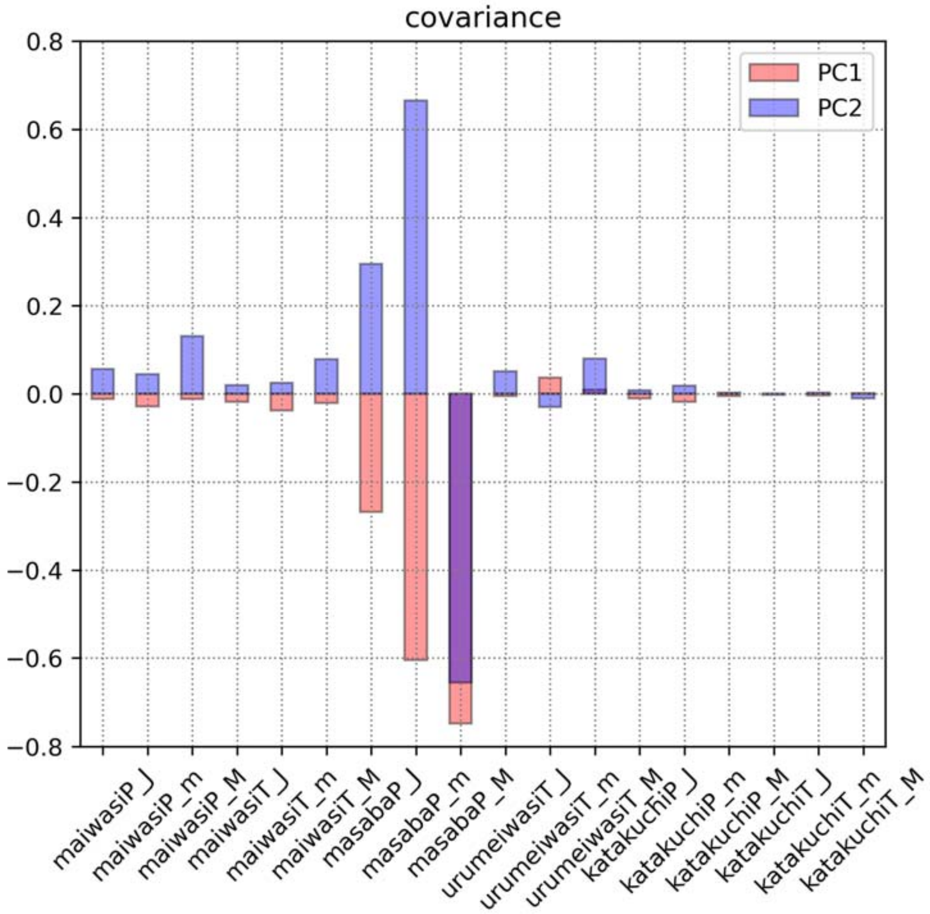

Figure 2. Eigen vector of PC1 (red) and PC2 (blue). Only “masaba” (chub mackerel) shows large amplitude. This is because now we are using wet weight data (raw data) and masaba has large value then the variation is also large.

Read out score of each PC from pca by the following command.

```
score = pd.DataFrame(pca.transform(data), index=sequence)
```

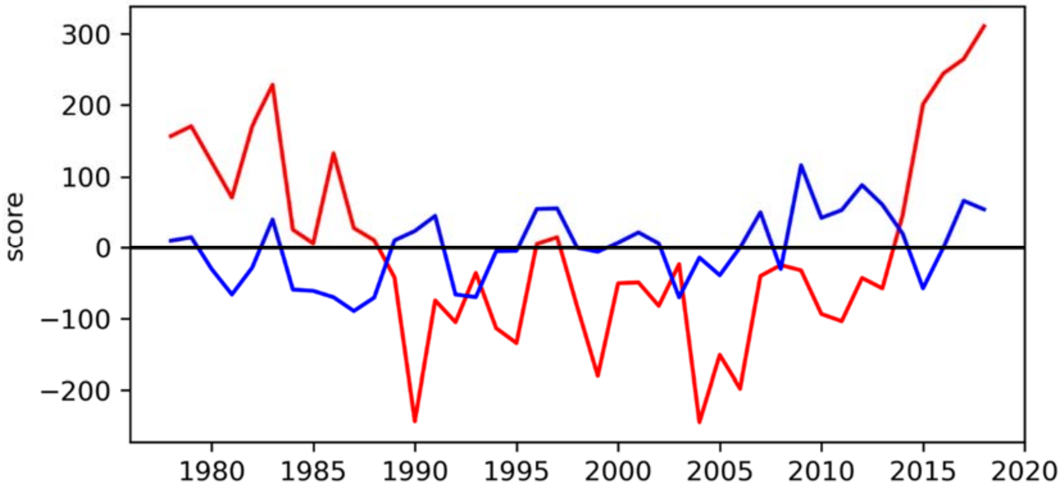

Figure 3. Scores of PC1 (red) and PC2 (blue).

From Figures 2 and 3, it is considered that PC1 represents weight decrease (increase) of chub mackerel during 1988-2014 (1978-1988 and 2014-2018). PC2 represents opposite response of juvenile & immature with mature of chub mackerel. The time scale is about 5 years.

PC1 eigen vector and PC2 eigen vector scatter plot will give you the characteristic of the response for each species.

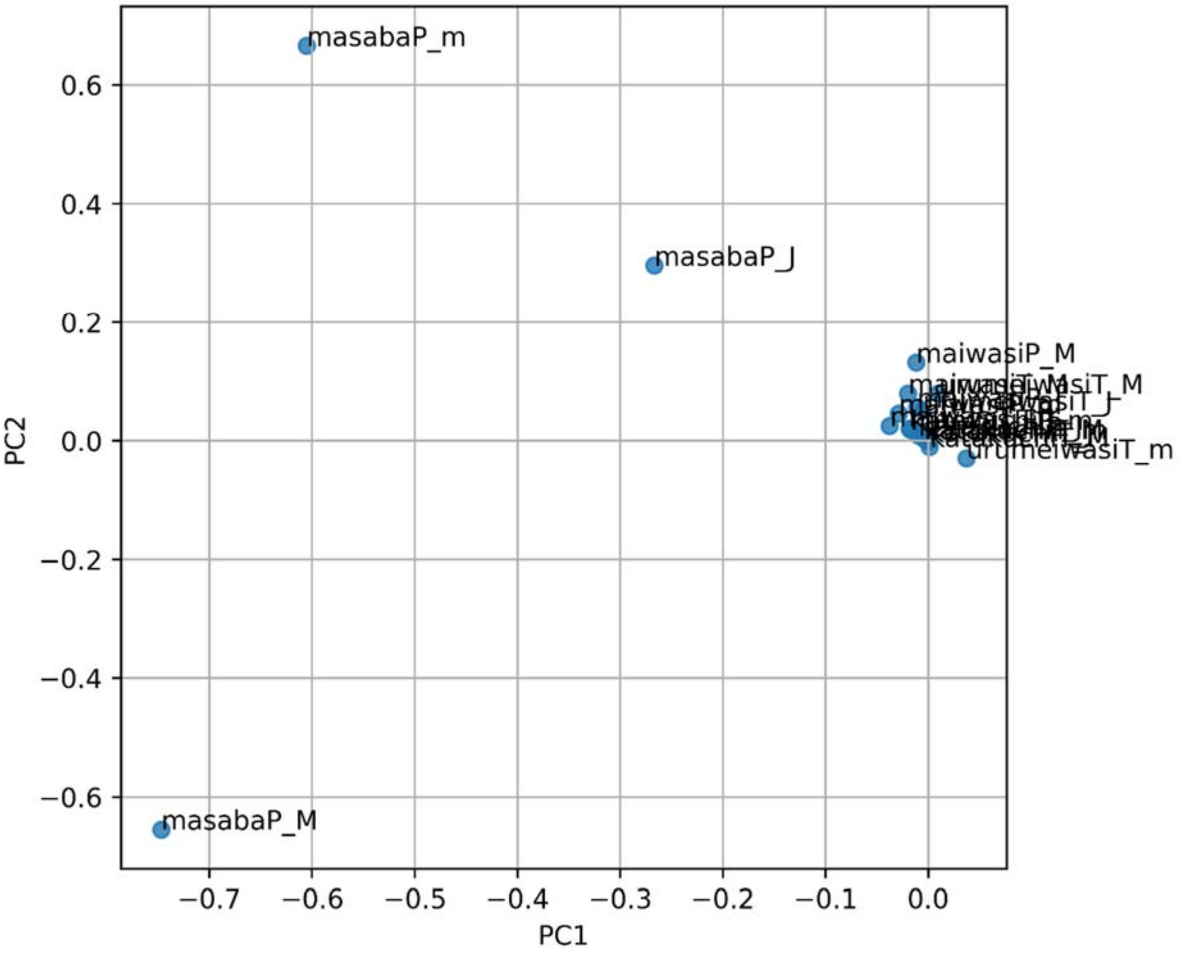

Figure 4. Scatter plot between PC1 and PC2 eigen vectors. Almost all species except for chub mackerel are at (0,0) which means its variation is small. Juvenile and immature chub mackerel will show similar variability regarding PC1 and PC2, but mature chub mackerel shows different feature.

2. Please conduct PCA based on correlation matrix (using standardized wet weight data).

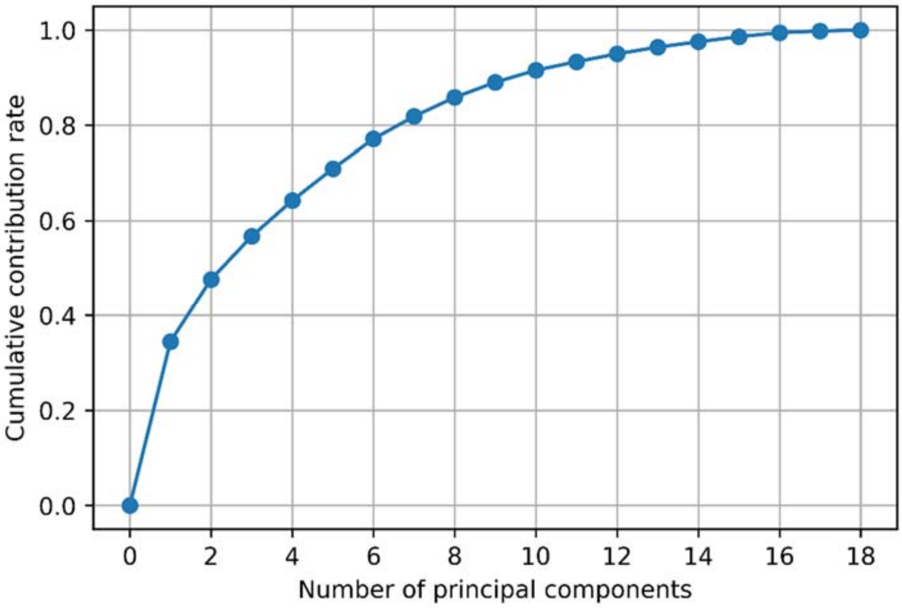

Figure 5. Cumulative contribution rate of PCs for the correlation matrix case. In this case, PC1 explained only about 35% of the total variance.

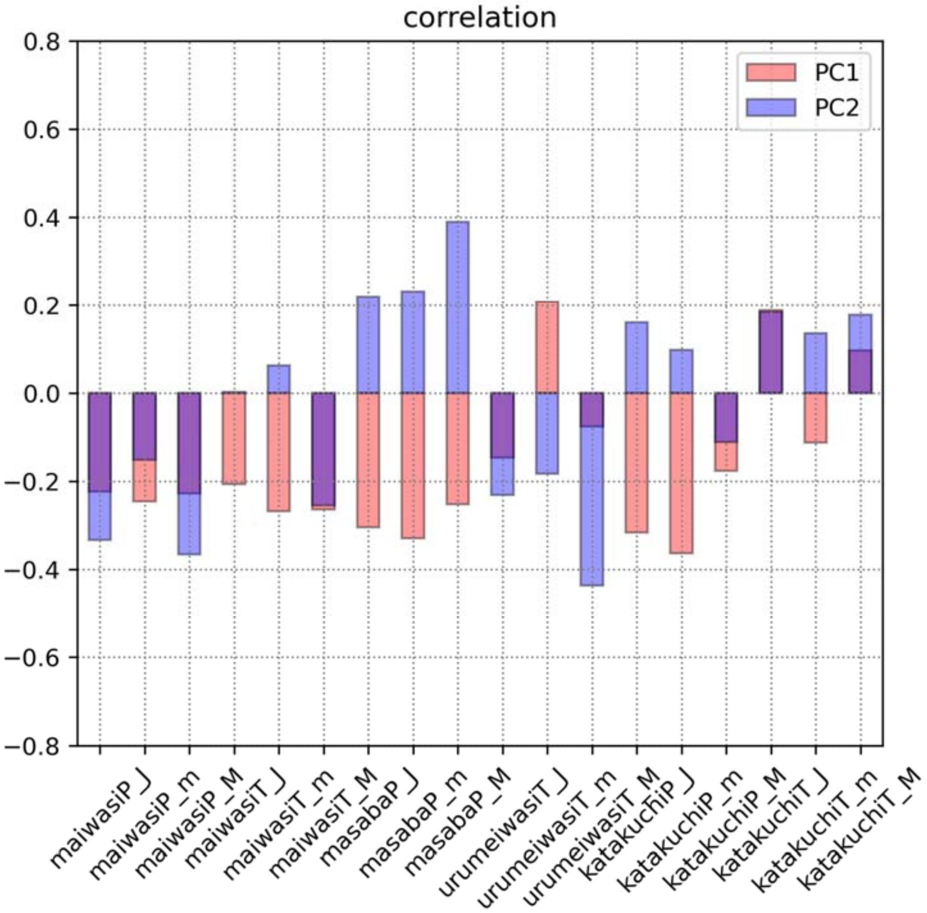

Figure 6. Eigen vector of PC1 (red) and PC2 (blue) for the correlation matrix case. Now we are using standardized data, then all species have larger amplitude.

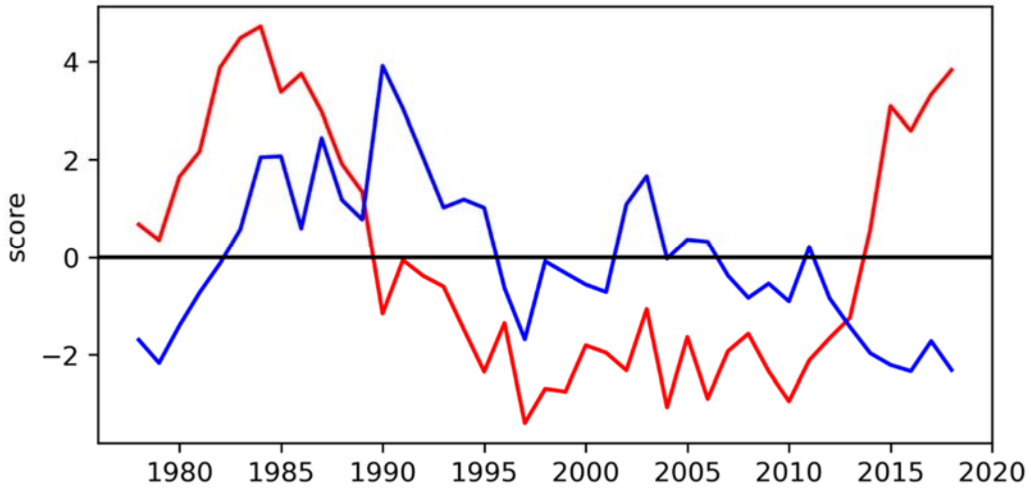

Figure 7. Scores of PC1 (red) and PC2 (blue) for the correlation matrix case.

From Figures 6 and 7, it is considered that PC1 represents see-saw weight variation between “maiwashi_P” & “urumeiwashi_T” and “masaba_P” & “katakuchi_P” & “katakuchi_T”. During 1980 it is well known that the biomass of “maiwashi_P” was high. The corresponding period, PC1 represents decreasing tendency of wet weight of almost all species except for “urumeiwashi_m”, ”katakuchi_T_J”, and “katakuchi_T_M”.

PC2 represents decresing (increasing) tendency of wet weight of “maiwashi_P” & “urumeiwashi_T” (“masaba_P” & “katakuchi_P” & “katakuchi_T”) during 1982-1995.

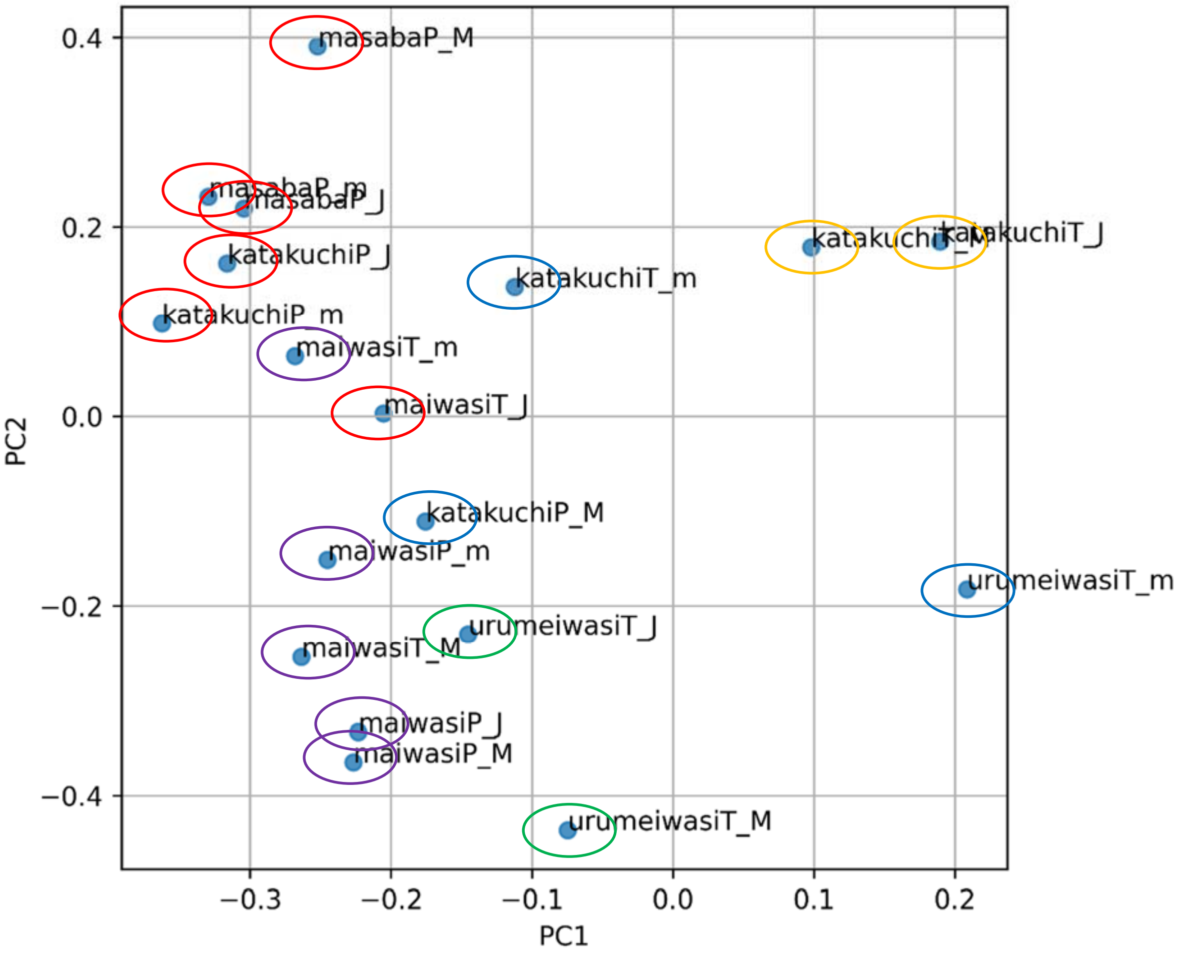

Figure 8. Scatter plot between PC1 and PC2 eigen vectors for the correlation matrix case. The colors show clusters by the complete methods. The cumulative contribution of PC1 and PC2 was less than 50% in this case. Therefore, PC1 and PC2 cannot completely capture the clustering features, but clusters except for blue were distributed separately on the PC1-PC2 plane.

The purple cluster locates in negative PC1 and negative PC2 (Figure 8), which results in large decrease in 1983-1990 (Figure 7).
The purple cluster locates in negative PC1 and positive PC2 (Figure 8), which results in earlier decrease (1978-1985) and recovery (after 1990) of wet weight (Figure 7).
The yellow cluster locates in positive PC1 and positive PC2 (Figure 8), which results in increase of wet weight during 1982-1992 (Figure 7).

The choice of covariance or correlation matrix depends on the objective of the study. For large scale analyses, correlation matrix is frequently used to detect the dominant variation pattern.