第 1 回 レポート課題

課題 1． Wine データの PCA

https://archive.ics.uci.edu/ml/datasets/Wine

上記の URL から wine.data，wine.names をダウンロードする．

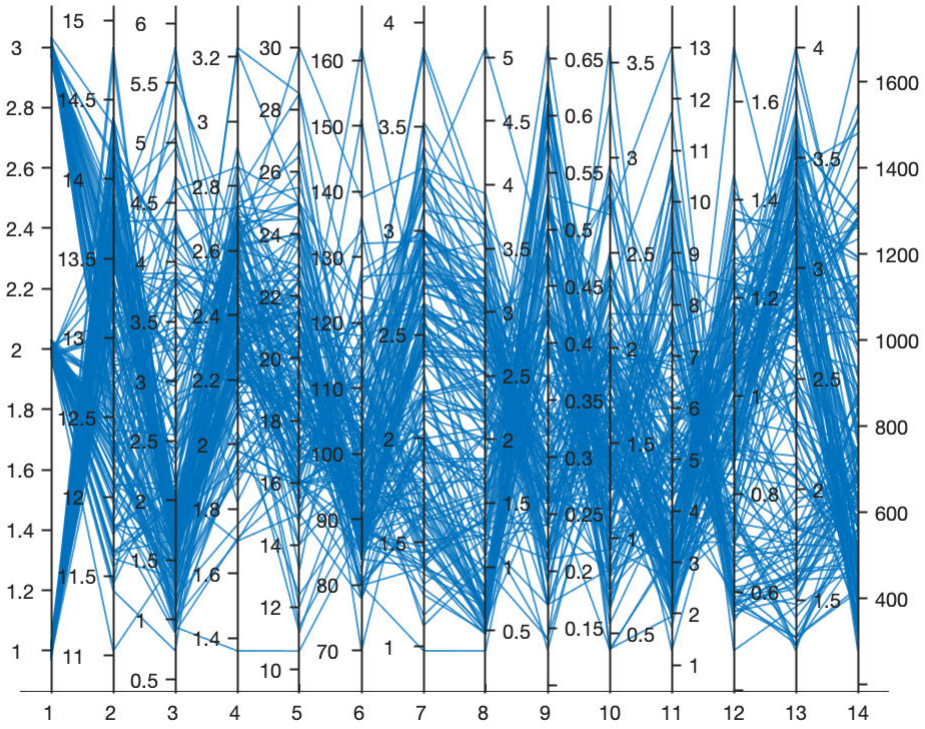
wine.csv をダブルクリックして数値行列として，インポートする．

ワークスペースに数値行列(178*14)の wine があることを確認する．

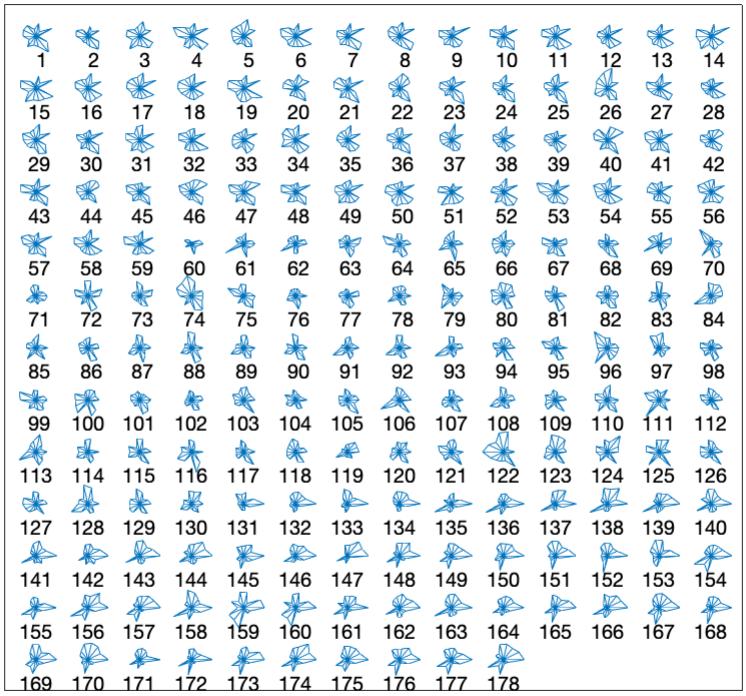
```
wine
```

```
wine = 178×14
10^3 ×
    0.0010    0.0142    0.0017    0.0024    0.0156    0.1270    0.0028    0.0031 ···
    0.0010    0.0132    0.0018    0.0021    0.0112    0.1000    0.0027    0.0028
    0.0010    0.0132    0.0024    0.0027    0.0186    0.1010    0.0028    0.0032
    0.0010    0.0144    0.0019    0.0025    0.0168    0.1130    0.0039    0.0035
    0.0010    0.0132    0.0026    0.0029    0.0210    0.1180    0.0028    0.0027
    0.0010    0.0142    0.0018    0.0025    0.0152    0.1120    0.0033    0.0034
    0.0010    0.0144    0.0019    0.0025    0.0146    0.0960    0.0025    0.0025
    0.0010    0.0141    0.0022    0.0026    0.0176    0.1210    0.0026    0.0025
    0.0010    0.0148    0.0016    0.0022    0.0140    0.0970    0.0028    0.0030
    0.0010    0.0139    0.0014    0.0023    0.0160    0.0980    0.0030    0.0032
     ⋮
```

```
parallelplot(wine)
```



```
glyphplot(wine)
```



・Wine データセットは，１列目を除いた２列目以降に，以下の１３個の特徴量を持つ.

1) Alcohol

2) Malic acid

~

13)Proline

これらの特徴量を標準化したものを x とする.

・１列目には，３つのクラスが格納されている.

class 1 59

class 2 71

class 3 48

これらのクラスラベルを y に格納する.

```
x = normalize(wine(:,2:end));
y = wine(:,1);
```

・pca の返り値

coeff ： 13*13 次元の主成分係数の行列．各列に 1 主成分の係数が含まれている。１列目・第 1 主成分の係数．2 列目・第 2 主成分の係数…となっている.

score ： 178*13 次元の主成分スコアの行列．主成分スコアは，主成分空間内の X の表現である．行は観測値に対応し，列は変数に対応している.

latent ： 13*1 次元のベクトル．上から第 1 主成分の寄与率，第 2 主成分の寄与率…となる.

cumsum(latent)/sum(latent) ：　累積寄与率．１に近いほど，データを説明できている.

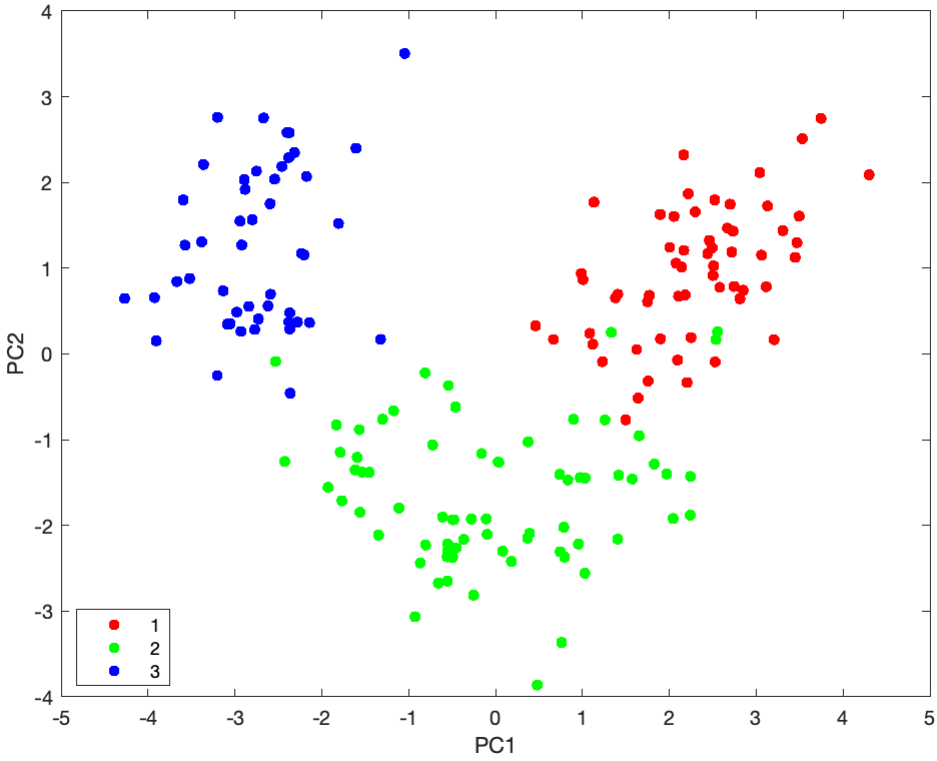
mu ： 1*13 次元の主成分分析の平均

```
[coeff, score, latent, ~, explained, mu] = pca(x);
```

分散が最大となる第一主成分と第二主成分でプロットする.

```
figure;
gscatter(score(:,1), score(:,2), y, 'rgb');
```

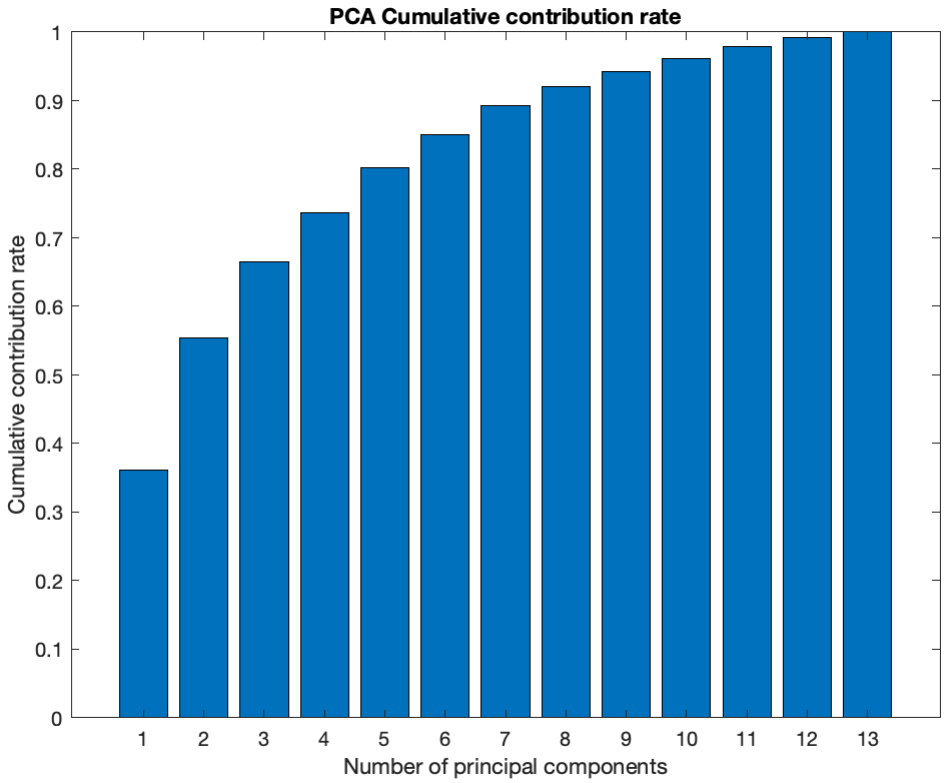
```
xlabel('PC1');
ylabel('PC2');
```



第一主成分と第二主成分で３クラスに，上手く分割していることがわかる.

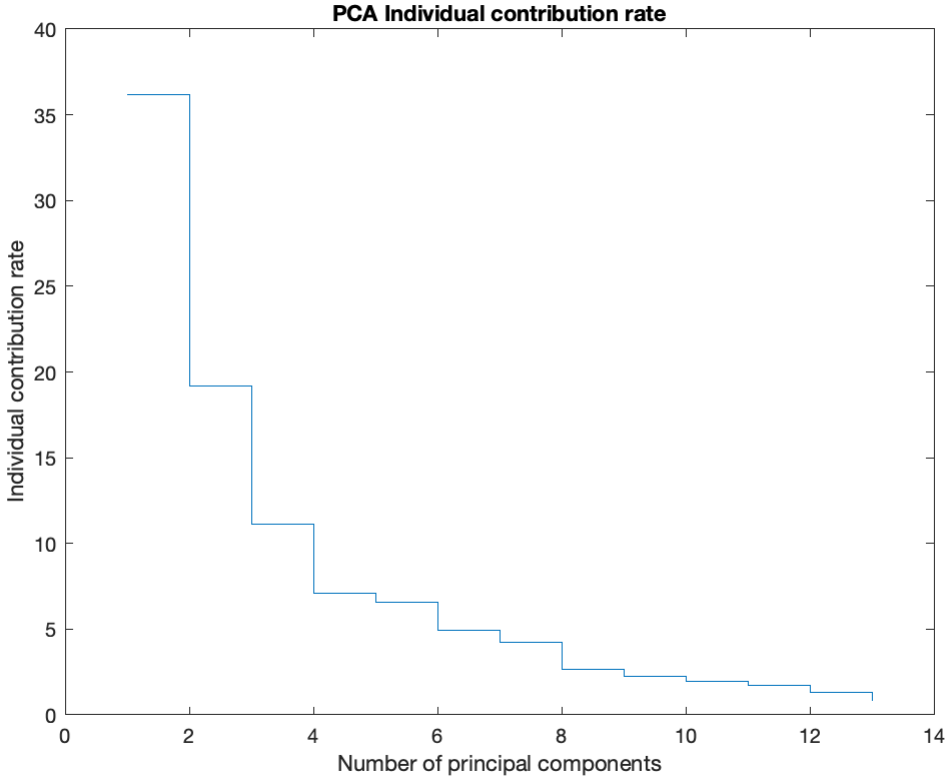
次に，累積寄与率を視覚化してみる.

```
% k = 1:13;
figure
bar(1:13, cumsum(latent)/sum(latent))
xlabel('Number of principal components')
ylabel('Cumulative contribution rate')
title('PCA Cumulative contribution rate')
```



第一主成分と第二主成分のみで，分散の約５５%を占めていることがわかる.

```
stairs(1:13, explained)
xlabel('Number of principal components')
ylabel('Individual contribution rate')
title('PCA Individual contribution rate')
```



第一主成分のみで約３５％，第二主成分のみで約２０％近くの分散を占めることがわかる．

以上の分析は，PCA により，１３次元の Wine データを２次元に次元削減が可能であることを示唆する．

課題２． 授業の感想

　演習という形で講義を進めていくのは，実践的で，自身の研究にも利用できるという点で良さを感じる．C++や Python などの言語を研究でよく使うが，MATLAB も割とライブラリが充実していて，使いこなせるようになると研究の幅はひろがると思う．しかし，ユーザー数があまり多くないため，ネット上の記事や情報量は少なめの言語であることが，デメリットの一つだと思う．MATLAB 自体を使うのは約３年ぶりで，ライブエディタの使い方や忘れていたコマンドを再び思い出せて，助かった．

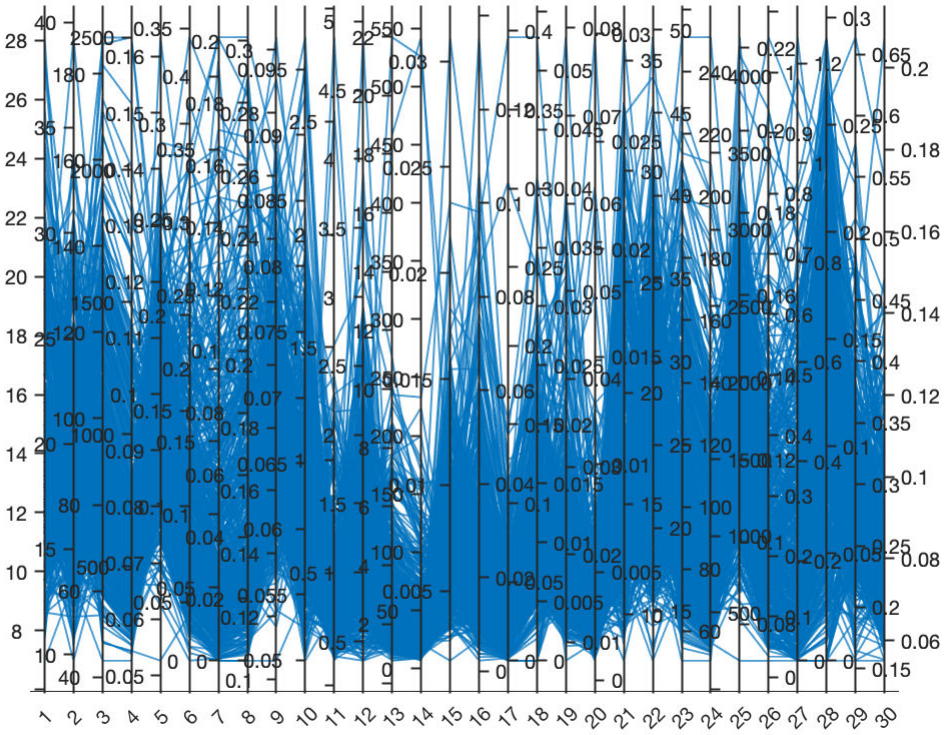
課題３． 自身で用意したデータセットでの PCA

https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

上記 URL の乳がんのデータセットを使う．

５６９人の被験者に対して，１列目に ID，２列目に診断結果(良性:B/悪性:M)，３列目以降は様々な計測値が格納されている．

```
data = readmatrix('data.csv');
data(:,3:end);
parallelplot(data(:,3:end))
```



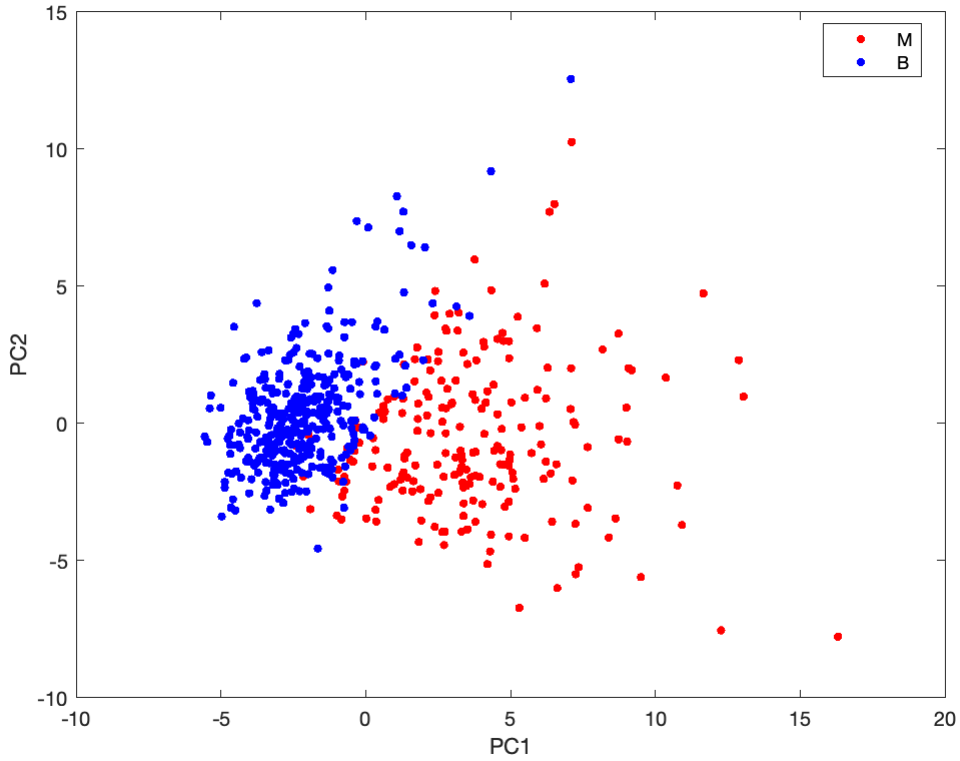
```
s = readtable('data.csv');
y = string(s{:,2})
```

```
y = 569×1 string
"M"
"M"
"M"
"M"
"M"
"M"
"M"
"M"
"M"
"M"
    ⋮
```

```
x = normalize(data(:,3:end))
```

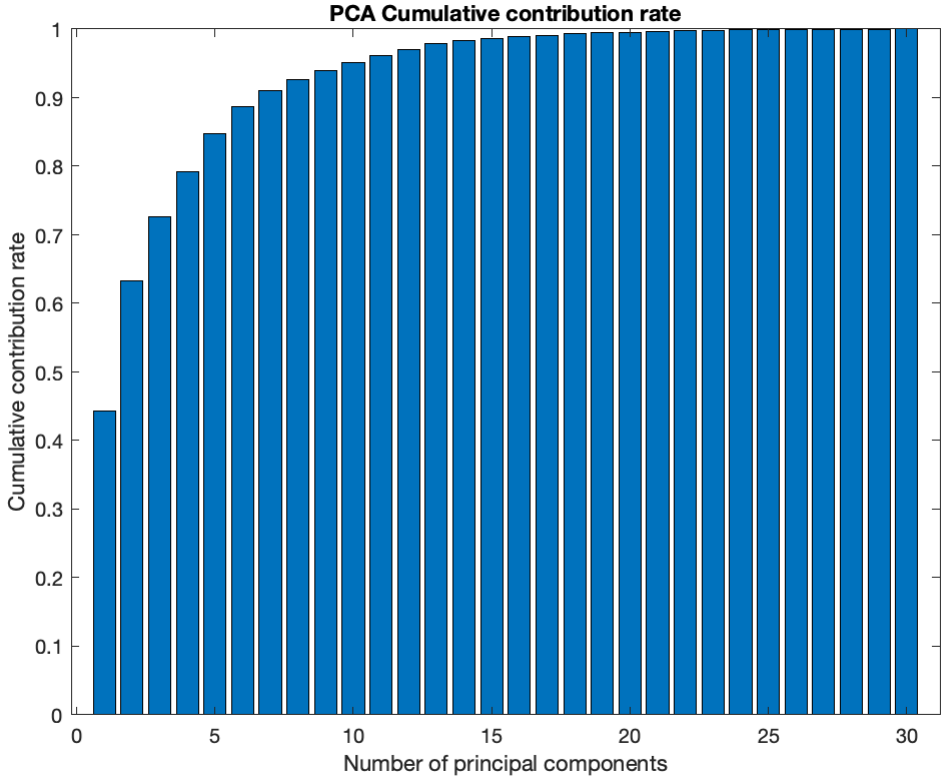
```
x = 569×30
    1.0961   -2.0715    1.2688    0.9835    1.5671    3.2806    2.6505    2.5302 ···
    1.8282   -0.3533    1.6845    1.9070   -0.8262   -0.4866   -0.0238    0.5477
    1.5785    0.4558    1.5651    1.5575    0.9414    1.0520    1.3623    2.0354
   -0.7682    0.2535   -0.5922   -0.7638    3.2807    3.3999    1.9142    1.4504
    1.7488   -1.1508    1.7750    1.8246    0.2801    0.5389    1.3698    1.4272
   -0.4760   -0.8346   -0.3868   -0.5052    2.2355    1.2432    0.8655    0.8239
    1.1699    0.1605    1.1371    1.0943   -0.1230    0.0882    0.2998    0.6464
   -0.1184    0.3581   -0.0728   -0.2188    1.6026    1.1391    0.0610    0.2817
   -0.3199    0.5883   -0.1839   -0.3839    2.1999    1.6825    1.2180    1.1497
   -0.4731    1.1045   -0.3292   -0.5086    1.5813    2.5611    1.7373    0.9409
    ⋮
```

```
[coeff, score, latent, ~, explained, mu] = pca(x);
figure;
gscatter(score(:,1), score(:,2), y, 'rb');
xlabel('PC1');
ylabel('PC2');
```



第一主成分と第二主成分のみで良性・悪性腫瘍の２クラスに，上手く分割していることがわかる.

```
figure
bar(1:30, cumsum(latent)/sum(latent))
xlabel('Number of principal components')
ylabel('Cumulative contribution rate')
title('PCA Cumulative contribution rate')
```



累積寄与率を見ると，第一主成分と第二主成分のみで，分散の約６０%以上を占めていることがわかる．また，第四主成分まで含めた場合，分散の約８０%以上を占めることがわかる.