正準判別分析

青木繁伸

1　目的

正準判別分析を行う。

2　使用法

```
from candis import candis
candis(data, verbose=True)
```

2.1　引数

data	説明変数と群変数のみからなるデータフレーム（最後列が群変数）
verbose	必要最小限のプリント出力をする

2.2　戻り値の名前

"means"	全体と各群の変数ごとの平均値
"univariate"	単変量統計
"between ss"	群間平方和・積和行列
"within ss"	群内平方和・積和行列
"pooled cov"	プールされた分散・共分散
"pooled r"	プールされた相関係数
"eigenvalues"	固有値
"canonical corr.coef."	正準相関係数
"WilksLambda"	Wilks の λ
"std.coef"	標準化判別係数
"structure"	構造行列
"coef"	判別係数
"centroids"	各群の重心
"score"	正準判別得点
"p Bayes"	各群に属するベイズ確率 p
"p"	各群に属する確率
"classification"	判別結果
"result"	判別結果表
"correctRate"	正判別率
"vnames"	説明変数の名前のベクトル
"ngroup"	群の数

"nax"　　解の次元数（個数）

3 使用例

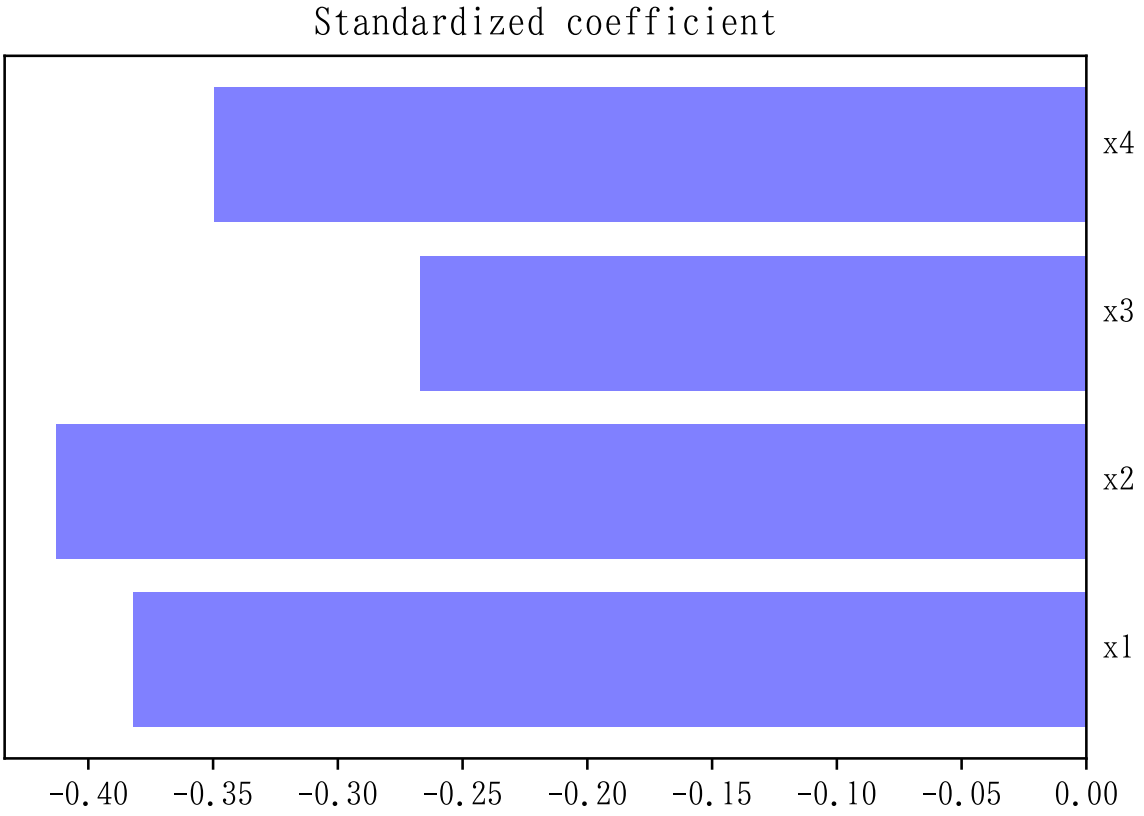
3.1 2 群判別

```
import scipy as sp
import pandas as pd

import sys
sys.path.append("statlib")
from gendat import gendat

data1 = pd.DataFrame(gendat(100, [0.23, 0.33, 0.45, 0.41, 0.25,
   0.37])*10+50)
data2 = pd.DataFrame(gendat(100, [0.23, 0.33, 0.45, 0.41, 0.25,
   0.37])*10+70)
data = pd.concat([data1, data2], axis=0)
data.columns = ["x"+str(i+1) for i in range(4)]
data["group"] = sp.repeat(["g1", "g2"], 100)

import sys
sys.path.append("statlib")
from candis import candis

a = candis(data, verbose=True)
```

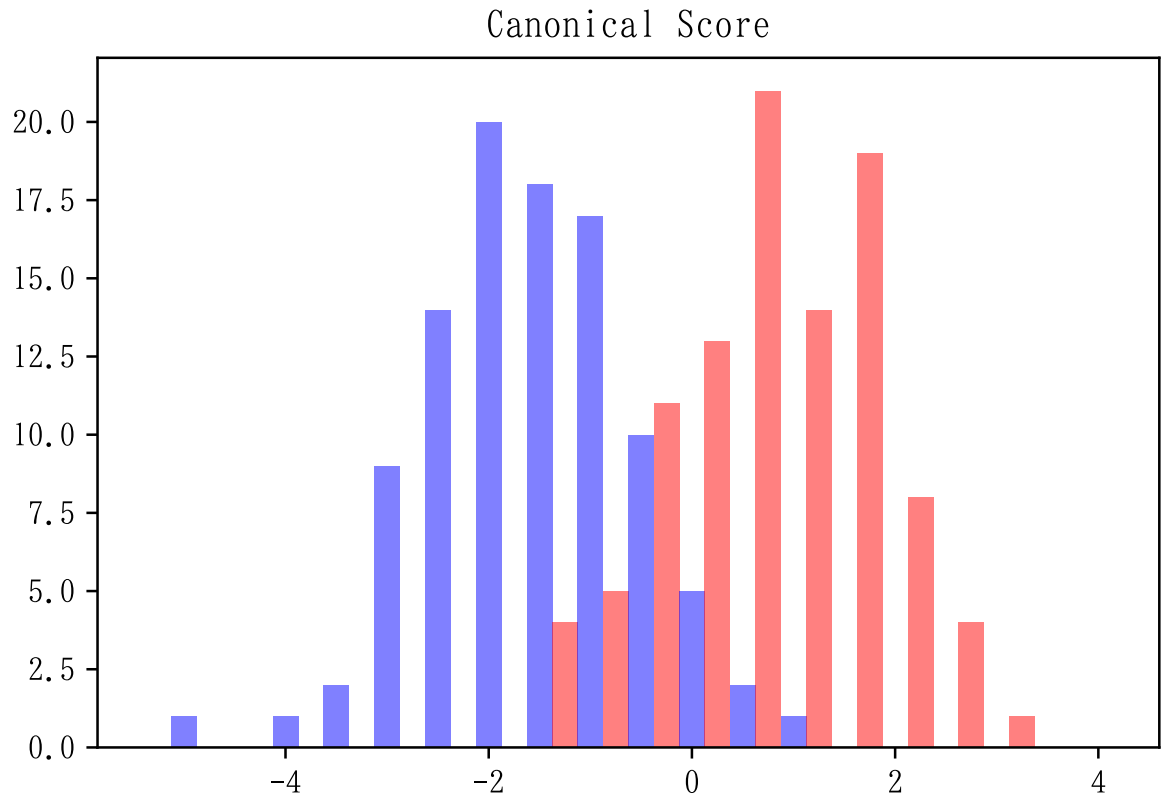
```
Wilks' Lambda
        Wilks' Lambda     chi.sq.  d.f.       p value
Axis 1       0.331903  216.170575     4  1.250035e-45

Discriminant coefficients
            Axis 1
x1       -0.038186
x2       -0.041290
x3       -0.026716
x4       -0.034975
constant  8.469983

Standardized discriminant coefficients
      Axis 1
x1 -0.381858
x2 -0.412897
x3 -0.267160
x4 -0.349749

Structure matrix
      Axis 1
```

```
x1 -0.708384
x2 -0.708384
x3 -0.708384
x4 -0.708384

Results of classification
    g1  g2
g1  91   9
g2   8  92
Correct rate = 91.5
```

```
from candis import candis_plot

candis_plot(a, type="std.coef")
```




```
from candis import candis_plot

candis_plot(a, type="score")
```



3.2 3 群以上の判別

```
import pandas as pd

data = pd.read_csv("data/iris.csv")

import sys
sys.path.append("statlib")
from candis import candis

a = candis(data, verbose=True)
```

```
Wilks' Lambda
        Wilks' Lambda     chi.sq.  d.f.        p value
Axis 1       0.023439  546.115296     8  8.870785e-113
Axis 2       0.777973   36.529664     3   5.786050e-08

Discriminant coefficients
             Axis 1    Axis 2
sl         0.829378  0.024102
sw         1.534473  2.164521
pl        -2.201212 -0.931921
pw        -2.810460  2.839188
constant   2.105106 -6.661473

Standardized discriminant coefficients
```

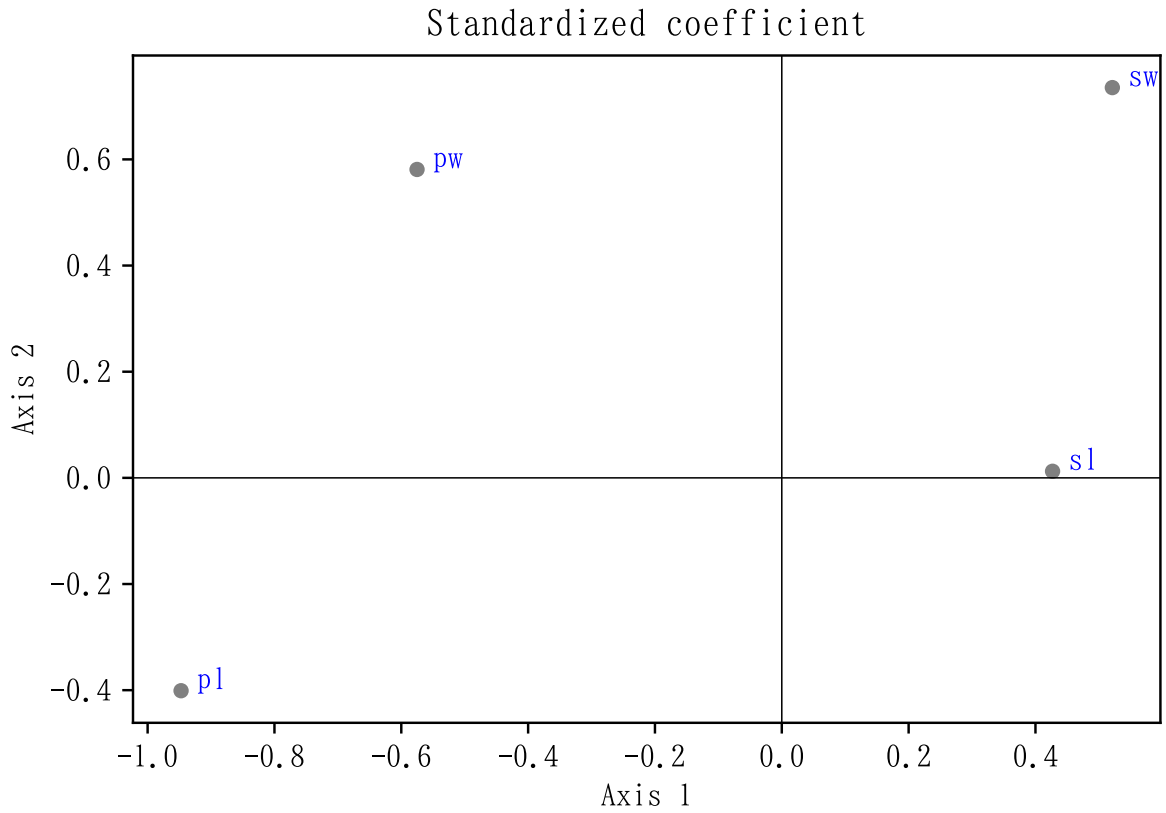
```
      Axis 1    Axis 2
sl  0.426955  0.012408
sw  0.521242  0.735261
pl -0.947257 -0.401038
pw -0.575161  0.581040

Structure matrix
      Axis 1    Axis 2
sl -0.222596  0.310812
sw  0.119012  0.863681
pl -0.706065  0.167701
pw -0.633178  0.737242

Results of classification
            setosa  versicolor  virginica
setosa          50           0          0
versicolor       0          48          2
virginica        0           1         49
Correct rate = 98.0
```

```
print(a["means"])
```

```
    grand mean  setosa  versicolor  virginica
sl    5.843333   5.006       5.936      6.588
sw    3.057333   3.428       2.770      2.974
pl    3.758000   1.462       4.260      5.552
pw    1.199333   0.246       1.326      2.026
```

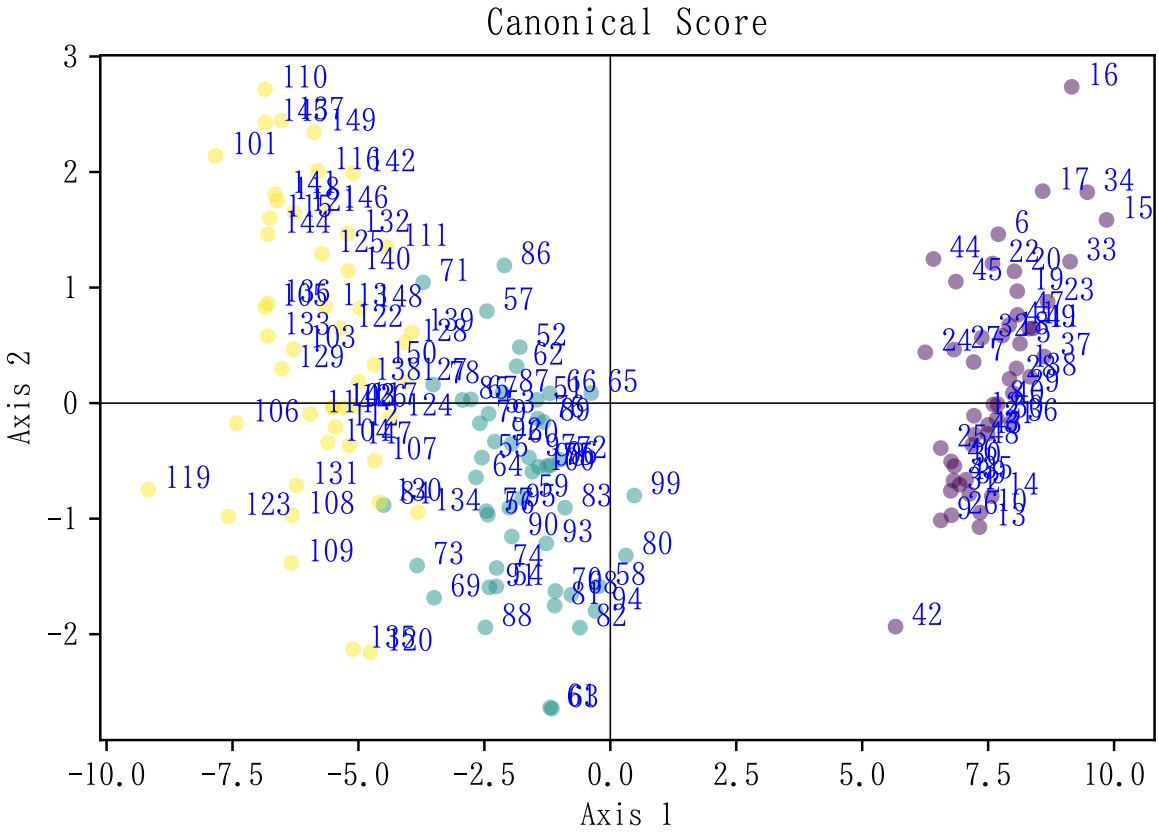
3.3 標準化判別係数

```
from candis import candis_plot

candis_plot(a, type="std.coef")
```



3.4 正準判別得点

```
from candis import candis_plot

candis_plot(a, type="score")
```


Canonical Score
Axis 2
Axis 1
3
2
1
0
-1
-2
-10.0
-7.5
-5.0
-2.5
0.0
2.5
5.0
7.5
10.0