正準判別分析

青木繁伸

2020 年 3 月 17 日

1 目的

正準判別分析を行う。

2 使用法

```
import sys
sys.path.append("statlib")
from multi import candis
candis(data, make_dummy=False, verbose=True)
```

結果のプロット

```
import sys
sys.path.append("statlib")
from multi import candis
candis_plot(obj, type="score", ax1=1, ax2=2, color="black", color2="blue", alpha=0.5)
```

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）
make_dummy	ダミー変数への変換が必要な場合には True を指定する。デフォルトは False
verbose	必要最小限のプリント出力をする
obj	candis() の返すオブジェクト
type	デフォルトは "score" で，正準判別得点をプロットする "stdcoef" で正準判別係数をプロットする。
ax1	横軸にとる解の番号
ax2	縦軸にとる解の番号
color	マークの色（デフォルトは黒）
color2	マークに添えるテキストの色（デフォルトは青）
alpha	アルファチャネル（デフォルトは 0.5）

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
"betweenss"	群間平方和・積和行列
"withinss"	群内平方和・積和行列
"pooledcov"	プールされた分散・共分散
"pooledr"	プールされた相関係数
"eigenvalues"	固有値
"canonicalcorr"	正準相関係数
"WilksLambda"	Wilks の λ
"stdcoef"	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
"pBayes"	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

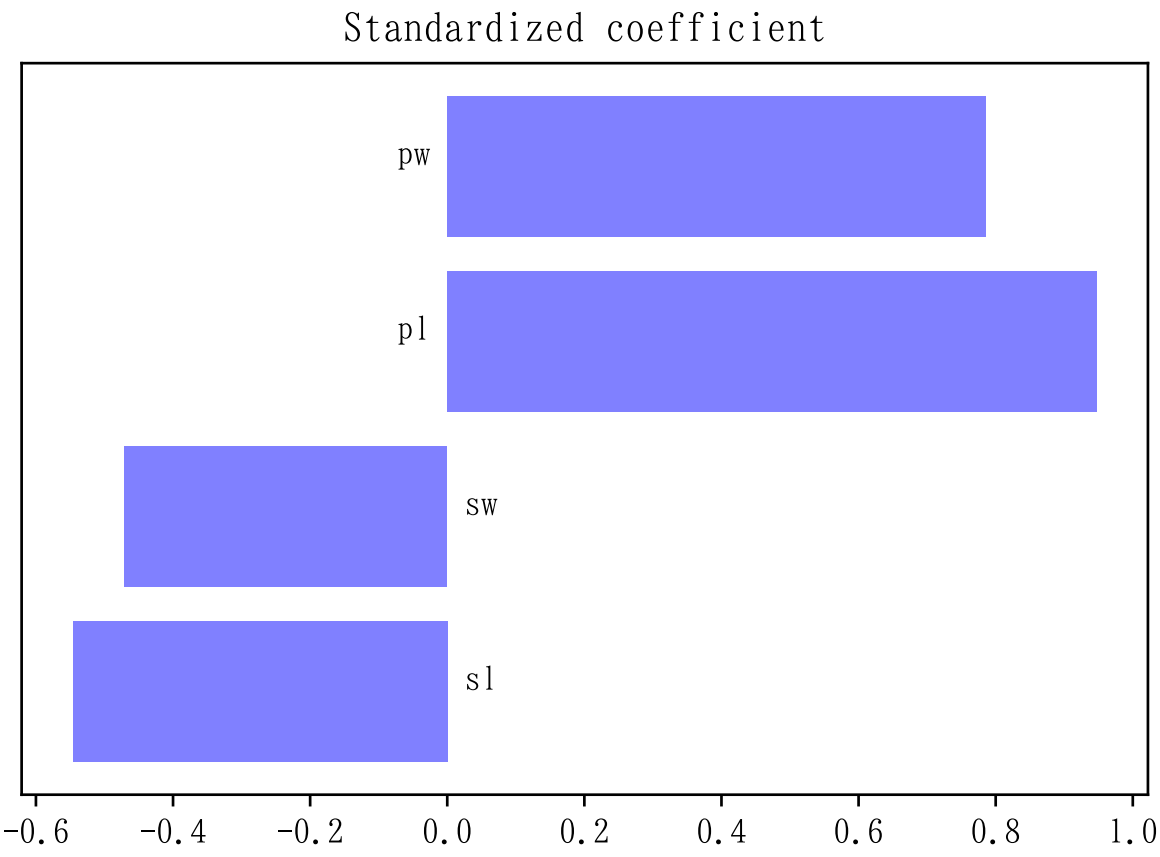
3.1 2 群判別

```
import pandas as pd

data = pd.read_csv("data/iris.csv")
data = data.iloc[50:, :]

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

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

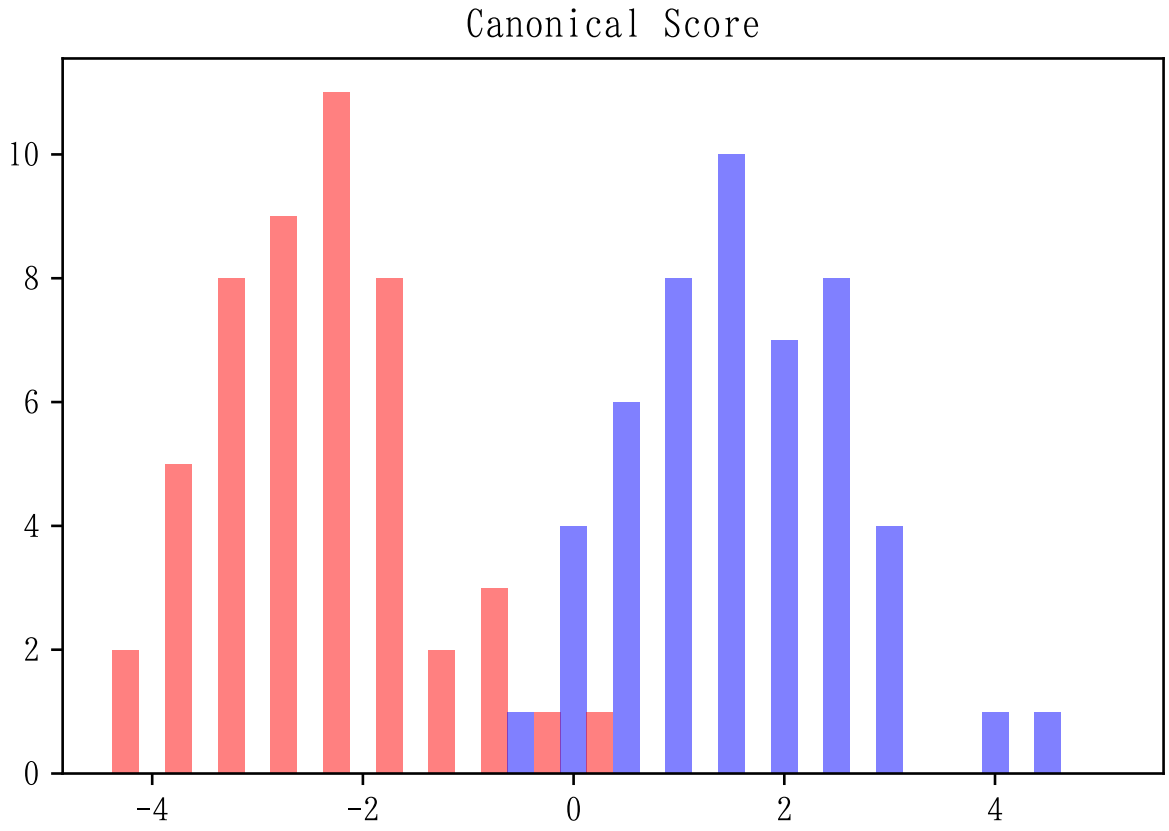
```
Wilks' Lambda
        Wilks' Lambda     chi.sq.  d.f.       p value
```

```
Axis 1        0.21611  147.068771     4  8.645274e-31

Discriminant coefficients
            Axis 1
sl       -0.943118
sw       -1.479429
pl        1.848451
pw        3.284730
constant -4.418986

Standardized discriminant coefficients
      Axis 1
sl -0.546185
sw -0.470720
pl  0.947414
pw  0.786070

Structure matrix
      Axis 1
sl  0.298567
sw  0.170031
pl  0.668495
pw  0.775718

Results of classification
            versicolor  virginica
versicolor          48          2
virginica            1         49
Correct rate = 97.0
```

```
from multi import candis_plot

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




```
from multi 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 multi 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
```

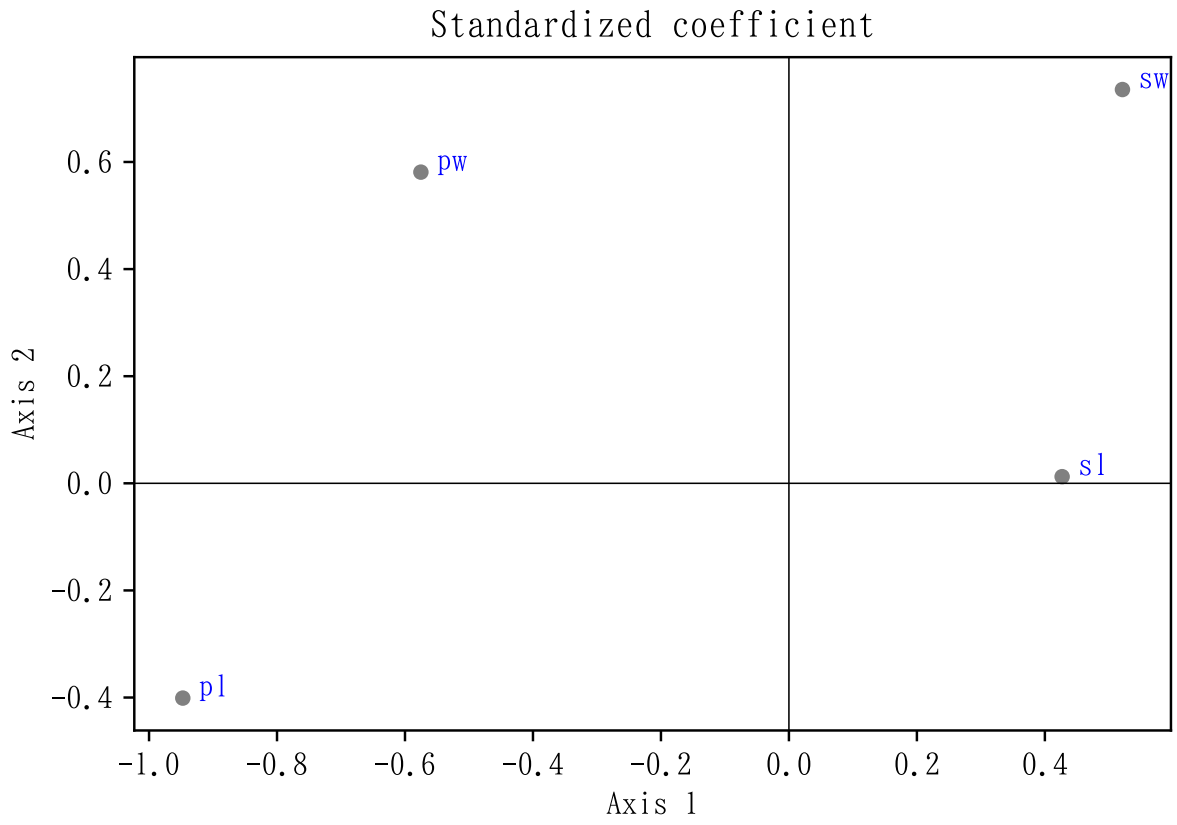
```
pw       -2.810460  2.839188
constant  2.105106 -6.661473

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
```

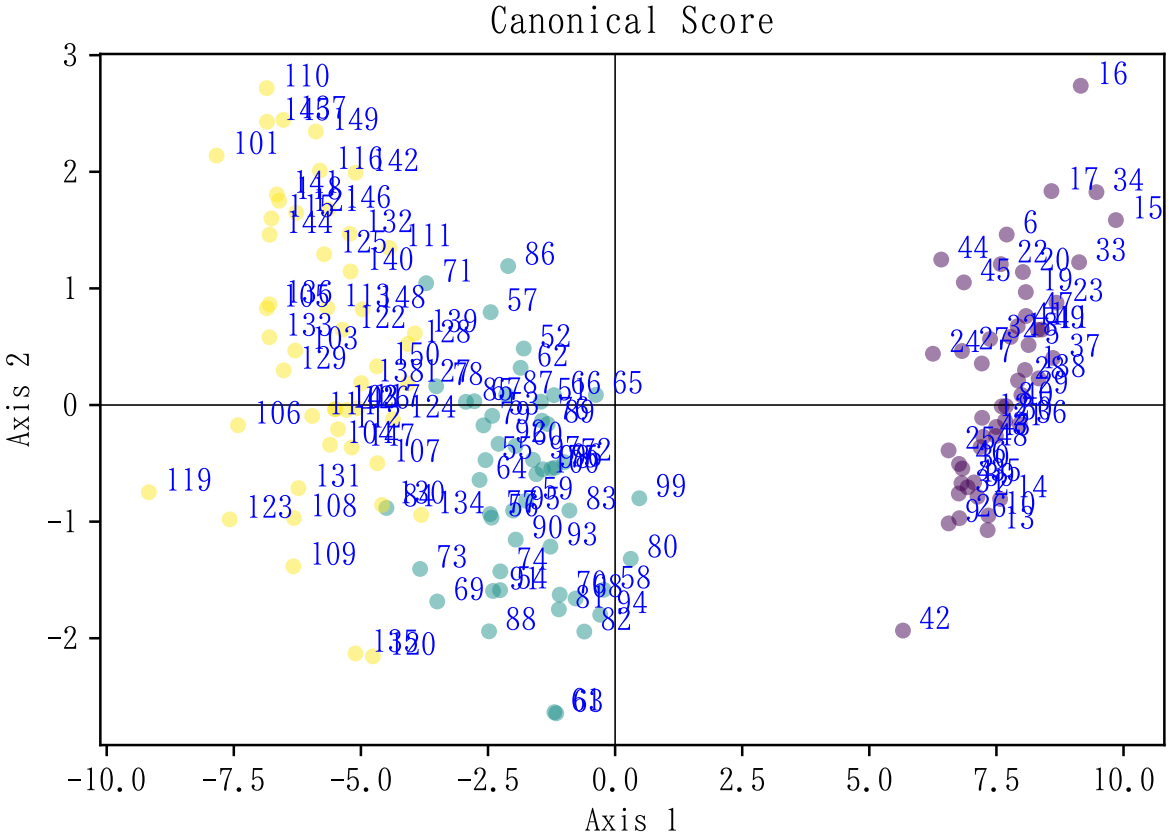
3.3 標準化判別係数

```
from multi import candis_plot

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



3.4 正準判別得点

```
from multi import candis_plot

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


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