

# 主成分分析その2，結果を解釈する

松田　史生[1]*・川瀬　雅也[2]

Aさん：じゃあC君がまだだけど，昨日の続きを始めましょっか．昨日はフィッシャーのアヤメのデータを使って主成分分析を復習しました［3種のアヤメ（*Iris setosa*，*I. virginica*，*I. versicolor*）各50個体の花弁（Petal）とがく（Sepal）それぞれの長さ（length）と幅（width）を計測したもの．単位はcm．4次元で総計150個体のデータ．“iris.csv”を生物工学会のHPからダウンロードして，C：直下のpydataに置く］．このうち花弁の長さと幅のデータだけを使って，主成分，主成分得点，寄与率，主成分負荷量などのおさらいをして……

B君：晩御飯食べに行ったなぁ．おいしかった．

Aさん：今日は，アヤメの全データを使って主成分分析をしてみましょう．

### 主成分分析の結果を解釈する

Aさん：Pythonでの作業は前回とほぼ同じなので，まとめたスクリプトを作ってみました．先輩，Spyderを使って打ち込んでみてください．前回と異なるのは，“iris”のデータフレームから，“data”を切り出すときに花弁とがくの長さと幅をすべて読み込んでいる点です．

```
*****pca1.pyのスクリプト
import pandas # pandasをインポート
import numpy
from matplotlib import pyplot # pyplotをインポート
from sklearn.decomposition import PCA
from scipy.stats import zscore
#irisのデータ読み込み
iris = pandas.read_csv("C:\pydata\iris.csv", sep=',')
data =
iris[["Petal.Length","Petal.Width","Sepal.Length",
"Sepal.Width"]]#データフレームで切り出し
species = iris["Species"]#リストとして切り出し
data_nomalized=zscore(data)#正規化
pca = PCA(n_components=4)#主成分分析準備
pca.fit(data_nomalized) #主成分分析実行
print("Contributions")#寄与率の表示
print(pca.explained_variance_ratio_)
print("Loadings")#主成分負荷量の計算と表示
loadings =
pca.components_*numpy.c_[numpy.sqrt(pca.explained_
variance_)]
print(loadings)
score = pca.fit_transform(data) #主成分スコアの計算とグラフの表示
pyplot.scatter(score[:, 0], score[:, 1])
for i in range(len(species)): #種名を書きだす
    pyplot.text(score[i, 0], score[i, 1], species[i])
******* pca1.pyここまで
```

Aさん：このスクリプトを“pca1.py”という名前でC：直下のpydataに保存しましょう

Bくん：保存までできたよ．

Aさん：次に，Spyderのツールバーにある緑の矢印を推すと，pca1.pyを実行することができます．うまくいくと主成分スコアプロットの結果が表示されるはずです（図1）．

```
In [15]: runfile('C:/pydata/pca1.py', wdir='C:/ pydata ')
Contributions
[0.72962445 0.22850762 0.03668922 0.00517871]
Loadings
[[ 0.99487699  0.96821173  0.89315091 -0.46168423]
```

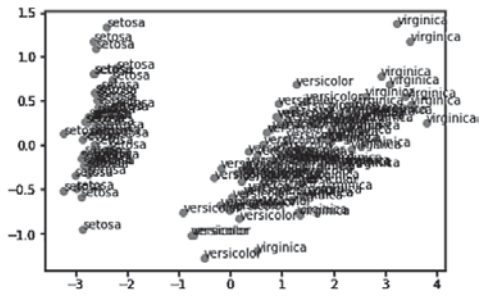


図1．pca1.pyの実行結果

**著者紹介**　[1]大阪大学大学院情報科学研究科（教授）　E-mail: fmatsuda@ist.osaka-u.ac.jp
[2]長浜バイオ大学（教授）

[ 0.02349363  0.06421425  0.3620387   0.88567345]
[ 0.05462939  0.24379667 -0.27658115  0.09393351]
[ 0.11573621 -0.07561196 -0.037732    0.01783586]]

B君：主成分スコアプロットの，種名の文字が重なり合って全然区別がつかないんだけど．

Aさん：先輩の机の上みたいですね．でもPythonは何でもできますから．あとで修正してみましょう．

**まずは寄与率**

Aさん：では，先輩，主成分分析の結果を解読するときに真っ先に注目すべき点は何でしたっけ？

B君：主成分スコアプロットでしょ．ぱっと見，第1主成分（横軸）の向きで大きく2つに分かれているし．

Aさん：先輩，まずは，寄与率ですよ．昨日の晩御飯の時にもC君とそう話してましたよね？

Contributions
[0.72962445 0.22850762 0.03668922 0.00517871]

が寄与率で，左から第1，第2主成分の寄与率です．

B君：第1主成分の寄与率が73 %，第2主成分の寄与率が23 %ということは，4次元のデータ全体のばらつきの96 %を第2主成分までで説明できているということ？

Aさん：そうですね．まず4次元のデータを2次元にうまく圧縮できた，わけですから，次元圧縮がうまくいったということになります．まずは累積寄与率を見て，次元圧縮がどのくらいうまくいったのかを確かめるのが大事ってことですね．

B君：でもさあ，第2主成分までで累積寄与率96 %はできすぎじゃないの？こんなにうまくいった例は論文でも見たことないよ．もっと多次元の多変量データを解析していて，第2主成分までの累積寄与率が30 %くらいになるのが，普通じゃないかな．

Aさん：確かにそうですね．どう考えたらいいんでしょ．でも，たとえば測定誤差のようなランダムなばらつきは圧縮できないので，誤差を多く含む多変量データだと，その分第1，第2主成分の寄与率が小さくなるんじゃないですか？

B君：このアヤメのデータセットは，定規で花弁やがくの長さを測ったものだから，計測誤差は小さいかも．それでも96 %はできすぎじゃない？

Aさん：第1主成分の寄与率が73 %だということは，このアヤメのデータセットのすべてのばらつきのうちの73 %が，第1主成分として要約された特徴（固有ベクトル）のばらつきで説明できるってことですよね．

B君：じゃあ，このアヤメのデータセットは，ばらつきを生む主な要因が1つだけあって，それに第2主成分を加えれば，ほとんど説明ができちゃうような結構単純なものだ，ということなのかな．

Aさん：となると，データセットの中にばらつきを生む要因が複数あるような場合も，相対的に第1，第2主成分の寄与率が小さくなりますね．第2主成分までの累積寄与率の大小だけでは，主成分分析の結果の良し悪しは決められないということですか．それから，たとえば第3とか第6主成分に重要な知見が含まれている場合がある，なんてこともあるのかしら．

B君：これまでそういう論文は，あまり見たことがないなぁ．大体，第1，第2主成分の2次元の主成分スコアプロットか，第3主成分までの3次元の主成分スコアプロットを図示して，ほらほらサンプルが分類できたで，主成分分析どや！という例が多いけど．

Aさん：こないだ読んだ論文もそうでしたね．たしかに．うーん．わかんなくなってきました．主成分分析ってそもそも何のためにするんでしたっけ？C君は次元圧縮法だって言ってたけど……

B君：そりゃあ，まずは，多変量データの特徴をパッと見た感じでつかみたいんだろ．多変量データのばらつきにもっとも寄与するベクトルが，第1主成分と第2主成分なんだったら，それを見たら，データの特性が一番つかみやすいからなんじゃないの？

Aさん：まずは多変量データを解析するときに知りたいのは，そういう大まかなデータの特徴ですよね．それがわからないと細かいところをうまく探せない．

B君：なにごともぱっと見が大事だって．

Aさん：（無視）．主成分の寄与率を見るべしというのは，どういうことなんでしょう？うーん．寄与率が大きいほど，その主成分が説明するばらつきが大きい，という意味ですよね．でも第1主成分の寄与率がたとえば15 %程度だったとしても，そのデータのばらつきにもっとも大きく寄与するベクトルであることは変わらないので，第1主成分を無視する理由は特にないってことですかね．

B君：だから主成分分析では，多変量データの特徴をパッとつかむために，まずは第1，第2主成分のスコアプ

ロットを示すんじゃない．全体に対する寄与がそれほど大きくなくても一番大事な特徴なのには変わりがないし．

Aさん：なるほどです．研究発表で示すスコアプロットに必ず，各主成分の寄与率を明記せよ，と指示されるのは，それがないと，第1，第2主成分で全体のばらつきのどれくらいが説明できているのかわからないからなんですね．主成分スコアプロットを見る前に，第1，第2主成分の重要性を把握しなさいということなんですね．

B君：でもさあ，第3とか第4主成分に意味がないかといわれたらそれも違うんじゃない？たとえば，小学生の身長，体重，年齢，足の速さとかのデータをたくさん集めて主成分分析をしたら，第1主成分はまず間違いなく，身長，体重，年齢を反映したベクトルになるはず．となると，もし成長度合い以外の個性のばらつきを議論したいなら，第2主成分以降に着目することが重要なんじゃないの？

Aさん：なるほど，まずは全体の傾向を調べたい，というときは，第1第2主成分が大事になるけど，主要なばらつきの傾向が最初からわかっていて，それとは無関係なばらつきを探したいなら，第3とか第4主成分に注目することもあり得ますよね．

### 次は主成分スコアプロット

B君：というわけで，横軸と縦軸の重要性がわかったから，次は主成分スコアプロットを見たいよね．

Aさん：pca1.pyの後半を書き換えましょう．スコアの計算以降を下記に変更してください．今回は，“score”に代入した主成分スコアプロットのデータのうち，1–50行目が*I. setosa*，51–100行目が*I. versicolor*と並んでいることを利用しました．各コマンドが何をしているのかは，自分で検索して調べてみてください．

```
score = pca.fit_transform(data) #スコアの計算
fig = pyplot.figure()#図を用意する
ax = fig.add_subplot(1,1,1)#subplot追加
#50個ずつsetosa, versicolor, virginicaの順
ax.scatter(score[0:50, 0], score[0:50, 1], c='red', label='setosa')
ax.scatter(score[50:100, 0], score[50:100, 1], c='blue', label='versicolor')
ax.scatter(score[100:150, 0], score[100:150, 1], c='green', label='virginica')

ax.set_title('PCA')
ax.set_xlabel('PC1 ')
ax.set_ylabel('PC2 ')
ax.legend(loc='upper left')
fig.show()
```

B君：わかりやすくなったじゃん（図2）．さっきも言ったけど，ぱっと見，第1主成分の向きで大きく2つに分かれてるでしょ．

Aさん：第1主成分の寄与率を確かめたあとだと，納得できちゃいますね．

B君：*I. setosa*の形態は他の2種にくらべて異なると言っていいはず．それから第1主成分の寄与率が73 %もあるんだから，*I. setosa*の形態は他の2種に大きく異なる，と言ってもいいんじゃないの？

Aさん：じゃあ，残りの*I. versicolor*と*I. verginica*の形態はそれに比べると類似する傾向がある，とも言えそうですね．なんとなくデータ全体の特徴がつかめてきましたね．あと，*I. versicolor*に比べ，*I. virginica*は

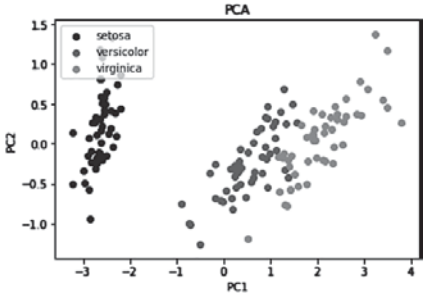


図2．修正後のpca1.pyの実行結果

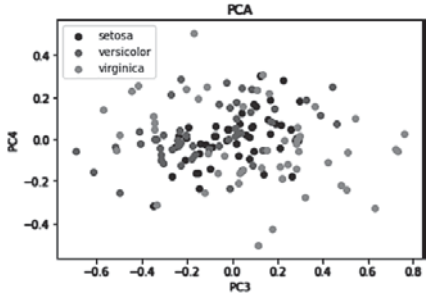


図3．第3，第4主成分のスコアプロット

第1主成分が大きく，第2主成分が小さくなる傾向がある，とも言えますかね.

B君：第2主成分までの累積寄与率が96 %もあるので，主成分スコアプロットから言えそうなことはこれだけというのはまあ，納得できるんだけど，じゃあ第3主成分や第4主成分には何も情報はないのかな.

Aさん：それじゃあ，確かめてみましょう．さっきのスクリプトをちょっと書き換えたらすぐできますよ.

B君:ほんとだ, はっきりした傾向はなさそうだね(図3). 2つ合わせて全体の7 %のばらつきしか反映していないし，測定の誤差と考えるといいと思います.

### ようやくローディング

Aさん：で，ようやく主成分負荷量（ローディング）の出番ですね．前回説明したように，正規化したデータセットで主成分分析をした時の第1主成分のローディングスコアとは，第1主成分スコアと元の変量（花弁の長さ，幅，がくの長さ，幅）との相関係数です．これを見ると，第1主成分がどのような特性を代表しているのかがわかるんでしたね（下記は見やすく整えてあります).

```
Loadings
#花弁の長さ，幅，がくの長さ，幅との相関係数
 [[ 0.994  0.968  0.893  -0.461] #第1主成分
 [ 0.023  0.064  0.362  0.885] #第2主成分
 [ 0.054  0.243  -0.276  0.093] #第3主成分
 [ 0.115  -0.075  -0.037  0.017]] #第4主成分
```

B君：第1主成分は，#がくの長さ，幅，花弁の長さとすごく強い正の相関がある．第1主成分は「花弁の長さ，幅，がくの長さ」を代表する軸であると言えるんじゃない. この辺は文学的, というか主観的になっちゃうんだけど.

Aさん：となると，*I. setosa*のクラスターは第1主成分得点が小さいので，他の品種にくらべて花弁の長さ，幅，がくの長さが小さい傾向があるってことですね．さらにこのばらつきが全体の76 %を説明するから，その傾向は他の傾向よりも大きい，となります．つぎに第2主成分は「がくの幅」を代表する軸である，と言えるわけですね.

B君：種内では，がくの幅に個体間でばらつきがある．あと，*I. virginica*は*I. versicolor*より，花弁の長さ，幅，がくの長さがやや大きく，さらに，がくの幅もやや大きい傾向がある，とも言えちゃうよね.

Aさん：でもそれ以上のことは，主成分分析からは言えないですよね．こないだ，2群3反復，総計6点の発現プロテオームデータの主成分分析についてX教授に相談したお返事をいただけない理由もわかりました. そもそも2群のデータの次元を圧縮する必要がないし, 寄与率を議論していない点もダメでした. ローディングプロットで端っこに来たタンパク質は，単に主成分との相関が高いだけなんだから，2群で有意に発現量差がある, なんて議論はできるはずがないのか. わーん．恥ずかしいな．X教授にメールでお詫びしよ.

B君：あと，主成分分析は多変量データの次元を圧縮して特徴を把握しやすくすることが目的だから，結果的に分類にも使えるけど，分類を目的とした方法ではないのか……そもそも第1，第2主成分までしか考慮していないし，新しいデータを追加して主成分分析をやり直したら, いろんなものがガラッと変わっちゃって, 異なる結果になってしまうかもしれないからね.

Aさん：何事でもそういうことはよく起こりますよ.

C君：Aさん，B先輩，こんにちは．遅くなってすみません．昨日晩御飯の後，本屋さんで面白い本を見つけちゃって，読み始めたら止まらなくって，寝坊しちゃいました．すみません．もう主成分分析の復習，っておわりました？