実践統計解析【第5回】

主成分分析その1，方法のおさらい

松田　史生[1]*・川瀬　雅也[2]

Aさん：今度のセミナーで紹介する論文に主成分分析が出てきたんですけど……

B君：オミクス解析の論文でも紹介するの？遺伝子発現とか代謝物蓄積プロファイルが似たサンプル同士の分類に使われているよね．あと，ローディングなんとかで端っこに来た遺伝子や代謝物に注目したりするんだったっけ．こないだの学会発表にもでてきたよ．なんかスマートでかっこいいじゃん．

Aさん：私もそれで，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を行ってみたんですよね．主成分プロットで2群が分かれたから2群のタンパク発現プロファイルには差があり，そのローディングプロットってやつで端っこに来た代謝物は，2群で有意に差がある，という議論をしてX教授にこれでいいかメールで質問したらその後お返事もらえないんです．

C君：先輩！主成分分析って，グループに分類する手法でも，2群間の優位差を調べる手法でもないんですよ．X教授，今頃あきれてものが言えないんじゃないかと思います．

B君：でもみんな主成分分析で，グループ化とかして発表してるじゃん．

Aさん：じゃ，主成分分析って何をするためのものなの？

主成分分析＝次元圧縮

C君：主成分分析は次元圧縮法の1つなんです．情報解析学演習で習いませんでしたっけ．

B君：……平成のころのことはあまり覚えてないなぁ….

Aさん：たしか，データを要約するイメージって先生が言っていたっけ？

C君：そうなんです．もしデータが2次元だったら簡単に図（散布図）にして傾向が議論できますよね．でも3次元になると途端に難しくなって，それ以上の多次元になると想像することすらできないじゃないですか．多次元のデータを要約するのが，次元圧縮法で，その手法の一つが主成分分析です．

B君：でも，主成分，主成分得点（スコア），寄与率，ローディングとか難しい言葉が出てきて，いつもわかんなくなるんだよなぁ．情報解析学演習の先生は，わかっているような大先生も，実はよくわかってなかったりすることもあるから，心配するな，って言ってたっけ．

C君：ちょうど僕も主成分分析の復習をしていたところなので，勉強しなおしてみましょうか？

Aさん：じゃあC君を先生役でやってみよう．

データの準備

C君：まず，フィッシャーのアヤメのデータを使いましょう．3種のアヤメ（*Iris setosa, I. virginica, I. versicolor*）について，各50個体の花弁（Petal）とがく（Sepal）の長さ（length）と幅（width）を計測したもの．単位はcmです．4次元で総計150個体のデータですね．“iris.csv”を生物工学会のHPからダウンロードして，いつも通り，C：直下のpydataに置いてあります．

```
In [67]: import pandas as pd # pandasをインポート
In [67]: iris = pd.read_csv("C:¥pydata¥iris.csv", sep=',')
```

で，データをirisというデータフレームに読み込みました．

```
In [67]: iris.head()
```

で確認すると，1列目がNo，2列目がSepal.Lengthで以下，Sepal.Width Petal.Length Petal.Width Speciesと並んでいます．最後はアヤメの種のデータです．

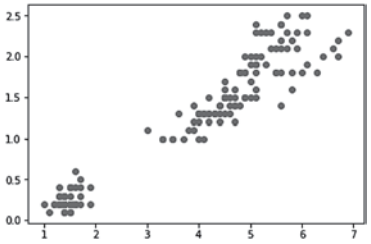
B君：4次元しかないの？

Aさん：それでも図に書けないんですよ．

C君：そこで，説明用に花弁（Petal）の長さと幅を2次元のデータを1次元に圧縮してみましょう．まずはここからですね．

Aさん：花弁の長さと幅を散布図にプロットしてみましょうか．

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）

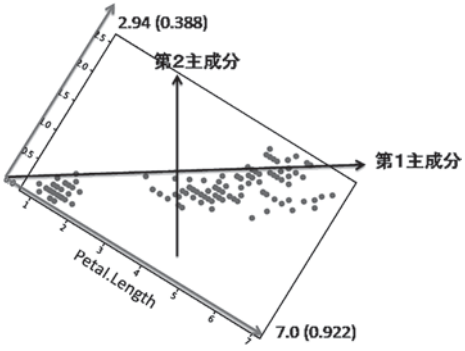


In [67]: from matplotlib import pyplot # pyplotのモジュールをインポート
In [67]: pyplot.scatter(iris["Petal.Length"], iris["Petal.Width"])
In [67]: import numpy as np #NumPyの読み込み
In [67]: np.corrcoef(iris["Petal.Length"], iris["Petal.Width"])
array([[1. , 0.96286543],
[0.96286543, 1.]])

C君：2変数の間の相関は0.962ですね.
B君：単に花弁の長さが大きいと幅も大きくなるってことなんじゃないの？

主成分，主成分得点（スコア）

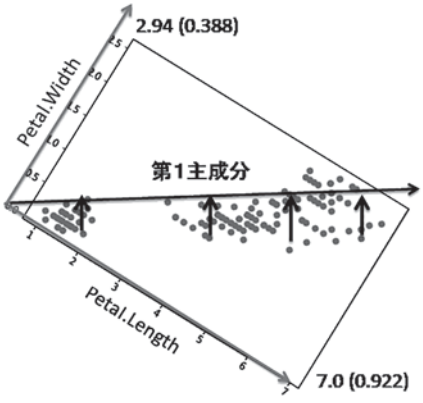
C君：それが次元圧縮のきっかけになるんです．2次元のデータを1次元に要約するというのは，データを説明するのに都合がいい新しい軸を1つ探す，ということです．主成分分析では，ばらつきがもっとも大きくなる軸を探します．たとえば，横軸方向の分散（ばらつき）がもっとも大きくなるようにグラフを回転しましょう．そしてその軸を第1主成分（principle component）とよぶことにします.



Aさん：では，主成分分析（PCA）をやってみましょう.

In [67]: from sklearn.decomposition import PCA#モジュール読み込み
In [67]: data = iris[["Petal.Length","Petal.Width"]]# データの切り出し
In [67]: pca = PCA(n_components=2)#主成分数は2
In [67]: pca.fit(data) #実行
In [67]: print(pca.components_)#主成分の表示
[[0.92177769 0.38771882]
[-0.38771882 0.92177769]]

C君:1行目の[0.92177769 0.38771882]が，第1主成分を，本来の軸の和として表現した固有ベクトルです．固有ベクトルの長さは1にするのがお作法ですね.
Aさん：なんで固有ベクトルっていうの？
C君：ええっとですね．ばらつきがもっとも大きくなる軸を探す，という問題をラグランジュの未定乗数法を使って解くんですが，その過程で出てくる式が変数の分散共分散行列の固有値を求める問題になっているからなんです．固有値が分散そのものであり，最大固有値をもつ固有ベクトルを第1主成分としている，わけです.
B君：その，ラグラン何とかは平成時代に聞いた気がする……．それで講義に置いていかれた気もする…….
Aさん：いろんなものに置いていかれたんですね….
C君:ではでは，各データの座標を主成分に投影しましょう．図に書くと第一主成分に各点を張り付けるイメージです．これを第1主成分得点（スコア）といいます．こうやって2次元データを1次元に圧縮するのが主成分分析です.



Aさん：fit_transform()関数で主成分得点が計算できるみたいです.

In [67]: score = pca.fit_transform(data) #スコアの計算

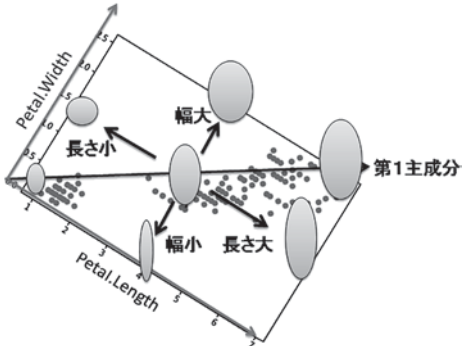
主成分の解釈は文学的

B君：で，この第1主成分得点の意味は何なの？ 1次元だから第1主成分得点が大きいか小さいかしかないんだけど．

C君：それにはですね，第1主成分の固有ベクトルを見ます．第1主成分得点は固有ベクトルを使って

第1主成分得点 = 0.922*Petal.length+0.388*Petal.width

と計算します．このときPetal.lengthが1 cm大きくなると，第1主成分は0.922増えます．Petal.widthが1 cm大きくなると第1主成分は0.388増えます．つまり，長さ，幅が大きい花弁は第1主成分得点が大きくなる．



といえます．さらにちょっと「文学的」に表現すると第1主成分得点とは「花弁の全体的な大きさ」の指標といえます．

B君：でも，幅が1 cm長くなるより，長さが1 cm大きくなる方が，寄与が大きくなる，ということは，第1主成分得点は花弁の幅よりも，長さの情報をより反映しているとも言えるんじゃない？

寄与率

Aさん：確かに固有ベクトルから意味を読み取るとしたら，「花弁の大きさ」というのは妥当な気もするけど，すこし乱暴かも．

C君：そこが，次元圧縮の難しいところみたいです．圧縮する以上, 元の情報の一部は失われちゃいますよね．

B君：じゃあ，どのくらいの情報が失われたわけ？

Aさん：確か寄与率か何かだっけ？

C君：そうです．データ全体ばらつきのうち，その主成分で説明できる割合です．

```
In [67]: #寄与率の表示
In [67]: print(pca.explained_variance_ratio_ )
[0.99025066 0.00974934]
```

第一主成分の寄与率が0.9902，第二主成分の寄与率が0.009なので，全体のばらつきの99％が第一主成分で説明できる，ってことです．

Aさん:ということは，この，花弁の幅と長さというデータセットの特徴の一つとして，最初にB先輩が行っていた「単に長さが大きいと幅も大きくなる」という関係がある，ということは言えるのかしら．

C君：むしろ，このデータセットばらつきをほとんど説明できているわけですから，代表的な特徴，あるいは傾向である，と言えると思います．

B君：そんなのわざわざ主成分分析をしなくたって，最初の相関係数を調べるだけでいいじゃん．

Aさん：でも，4次元とか多次元のデータになると相関係数だけからはわかりにくいですよね．主成分分析だとうまくデータの特徴が抽出できるかもしれないってことかな．

C君：それから，もう一つ．第1主成分得点が大きいサンプルは「花弁が幅も長さも大きい」傾向がある，ともいえます．

B君：でも，さっき，第1主成分得点は花弁の幅よりも，長さの情報をより反映しているというのはどう考えたらいいの？

Aさん：あ，でもそれは，花弁の長さのほうがばらつきが大きいからではないでしょうか？花弁の長さのデータは1–7 cmの間でばらついていますが，花弁の幅は0–2.5 cmの間です．

C君：あ，その点もかなり重要です．あとで主成分負荷量のところで説明します．

C君：続けて第2主成分です．第1主成分と直交するベクトルの中でもっとも分散が大きくなるものを，第2主成分とします．今回の例では2次元のデータなので，良い例ではありませんが，第2主成分の固有ベクトルは，[-0.38771882 0.92177769]です．

第2主成分得点 = -0.388*Petal.length+0.922*Petal.width

となります．第2主成分得点は花弁の長さが小さくなると大きくなり，花弁の幅が大きくなると大きくなります．つまり，第2主成分得点が大きいサンプルは，

花弁の長さが小さく，幅が大きいということで，いわば「花弁の扁平さ」の指標であるといえます

B君：これまた文学的だね．でも第2主成分の寄与率って1％もないんだろ．

Aさん：ということは，今回の花弁の長さと幅のデータセットの特徴として，花弁の扁平さのばらつきはかなり小さい，重要なファクターではないと言えるわけね．

C君：今回の例では元のデータが2次元なのでこれでおしまいですが，実際のn次元のデータでは，次に第1主成分，第2主成分に直交するベクトルの中で分散がもっとも大きくなるものを，第3主成分する，というような作業を第n主成分まで行っています（実際には分散共分散行列の固有ベクトルとしていっぺんに計算しています）．

主成分負荷量

C君：ではいよいよ主成分負荷量です．ローディングとも言います．まず，これまでは元の実測データと主成分との関係をわかりやすくするため，正規化抜きで主成分分析を行いました．が，実際に解析する際には，正規化を行うのが普通です．

B君：正規化ってなんだっけ？

Aさん：データの平均をゼロ，分散が1になるように変換するんですよね．今読んでいる論文ではZスコア化と呼んでいます．

B君：でもさあ，このフィッシャーのアヤメのデータってどうみても花弁の長さも幅も正規分布には従っていないように見えるけど，そんな変換しちゃっていいの？

C君：そこはずっと気になっていたのですが，主成分負荷量の値を評価するために，どうしても必要になるので，ここではスルーさせてください．主成分負荷量とは固有ベクトルと標準偏差のベクトルの積として計算します．

Aさん：計算してみましょ．

```
In [67]: from scipy.stats import zscore #zscoreをインポート
In [67]: data_nomalized=zscore(data)# データを標準化
In [67]: print(pca.components_)# 主成分の固有ベクトル
[[ 0.70710678  0.70710678]
 [-0.70710678  0.70710678]]
In [67]: #主成分負荷量を計算
In [67]: loadings = pca.components_*
np.c_[np.sqrt(pca.explained_variance_)]
In [67]: print(loadings)
 [[ 0.99399171  0.99399171]
 [-0.13671831  0.13671831]]
```

C君：一行目が第一主成分へのPetal.LengthとPetal.Widthの主成分負荷量です．花弁の長さの主成分負荷量は0.99ですが，これは花弁の長さ（正規化済）と第一主成分得点との間の相関係数なんです．花弁の長さが大きくなるほど第一主成分得点が大きくなる正の相関がある，ということです．また，相関係数は-1から1の間になり，1に近いほど正に相関する，ので主成分負荷量が0.99ということは，すごく相関が強いと言えます．さらに，花弁の幅の主成分負荷量も0.99なので，花弁の長さも幅も第一主成分得点と同じくらい正に強く相関している，と言えます．つまり，主成分負荷量の比較から第一主成分得点には花弁の長さも幅も同じだけ反映している，と言えるわけです．

Aさん：主成分負荷量が大きい変量ほど，主成分により大きく反映しているってことね．

B君：確かに，これだけ1に近いとそうかなって納得するけど，正規分布に従っていないデータで，相関係数の大小でそんな議論しちゃっていいの？

C君： 時々ハマるのは，場合によっては第1主成分の固有ベクトルが[0.922 0.388]ではなく，[-0.922 -0.388]となることがあります．向きが逆になっただけなんですが，第1主成分得点が「花弁の全体的な小ささ」の指標になっちゃうんですね．

Aさん：あと，標準化は注意が必要ってX教授がいっていたような．その辺は，また次回ってことで，晩御飯食べに行きましょ．