クレジット:
UTokyo Online Education 統計データ解析 Ⅱ 2018 小池祐太

東京大学 THE UNIVERSITY OF TOKYO

CC BY NC ND

統計データ解析 II (平成３０年度)

東京大学大学院数理科学研究科
統計データ解析教育研究グループ

村田　昇 (早稲田大学, 東京大学)
吉田朋広 (東京大学)
小池祐太 (東京大学)

第 6 章 主成分分析

6.1. 目的

主成分分析 *(principal component analysis)* とは, 多数の変数/データが与えられたとき, 変数/データたちのもつ情報を効率的に縮約して少数の特徴量を構成することで, 変数/データ間の関係を明らかにするための分析法である. 主成分分析では, 特徴量は変数/データたちの線形結合として構成する.

数式で表すと以下のようになる. 与えられた変数を $X_1, \dots, X_p$ としたとき, d を p 以下の正の整数とし (通常 p より小さくとる), $X_1, \dots, X_p$ の線形結合として表される d 個の変数

$$Z_k = a_{1k}X_1 + \cdots + a_{pk}X_p \qquad (k = 1, \dots, d)$$

を, もとの変数のもつ情報を最大限保持しつつ適切に構成することが目的である. ここで, Z_k と Z_k の (0 でない) 定数倍は互いに同じ情報量をもつので, そのような定数倍の任意性をなくすため, ベクトル $\boldsymbol{a}_k := (a_{1k}, \dots, a_{pk})^\top$ の長さが 1 となるようにする. すなわち,

$$\|\boldsymbol{a}_k\|^2 := \sum_{j=1}^{p} a_{jk}^2 = 1$$

と仮定する.

幾何学的にいうと, 観測データの含まれる p 次元空間にうまく座標軸を設定することにより, その座標軸上にデータのもつ情報が最大限反映されるようにすることが目的である.

6.2. 計算法

6.2.1. $d = 1$ の場合. 組 $(X_1, \dots, X_p)$ に対する n 個の観測データ $\{(x_{i1}, \dots, x_{ip})\}_{i=1}^n$ が与えられているとする. いま, i 番目の観測データに対応する p 次元ベクトルを $\boldsymbol{x}_i := (x_{i1}, \dots, x_{ip})^\top$ とすると, われわれの目的は, 長さ 1 の p 次元ベクトル $\boldsymbol{a} = (a_1, \dots, a_p)^\top$ をうまく選んで, 観測データ $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ のもつ情報を最大限保持するように 1 変量データ $\boldsymbol{a} \cdot \boldsymbol{x}_1, \dots, \boldsymbol{a} \cdot \boldsymbol{x}_n$ を構成することである. ところで, 各 $\boldsymbol{x}_i$ は p 次元空間内の点だとみなせるが, このとき $(\boldsymbol{a} \cdot \boldsymbol{x}_i)\boldsymbol{a}$ はベクトル $\boldsymbol{a}$ で張られる部分空間 (直線) への点 $\boldsymbol{x}_i$ の直交射影に一致する (図 1 参照). すなわち, $\boldsymbol{a} \cdot \boldsymbol{x}_i$ はベクトル $\boldsymbol{x}_i$ の $\boldsymbol{a}$-方向成分であると解釈できる. このような幾何学的考察から, ベクトル $\boldsymbol{a}$ の適切な選び方の指針としては以下の考え方が自然である:

- 構成した特徴量がもとのデータのばらつきを最大限反映するように, $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ たちのばらつきが最も大きい方向 $\boldsymbol{a}$ を選ぶ. すなわち,

  $$\sum_{i=1}^{n} (\boldsymbol{a} \cdot \boldsymbol{x}_i - \boldsymbol{a} \cdot \bar{\boldsymbol{x}})^2$$

  を最大化するように $\boldsymbol{a}$ を選ぶ. ここに,

  $$\bar{\boldsymbol{x}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i$$

  であり, 従って $\boldsymbol{a} \cdot \bar{\boldsymbol{x}}$ は $\boldsymbol{a} \cdot \boldsymbol{x}_1, \dots, \boldsymbol{a} \cdot \boldsymbol{x}_n$ の平均に対応する.

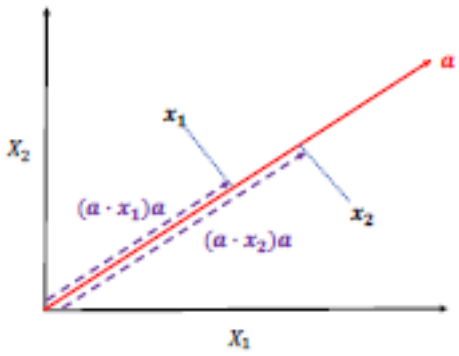


図 1. ベクトル $\boldsymbol{a}$ への観測データの直交射影 ($p=2, n=2$ の場合)

以上をまとめると, 関数

$$f(\boldsymbol{a}) = \sum_{i=1}^{n} (\boldsymbol{a}\cdot\boldsymbol{x}_i - \boldsymbol{a}\cdot\boldsymbol{x})^2$$

を制約条件 $\|\boldsymbol{a}\| = 1$ の下で最大化するように, ベクトル $\boldsymbol{a}$ を選ぶのが方針となる. 詳細な導出は 6.5 節を参照してほしいが, そのようなベクトル $\boldsymbol{a}$ は存在して次の性質をもつことがわかる:

- $f(\boldsymbol{a})$ は行列 $\boldsymbol{X}^\top\boldsymbol{X}$ の固有値であり, $\boldsymbol{a}$ はこの固有値に対する固有ベクトルである.

ただし, $n\times p$ 行列 $\boldsymbol{X}$ を

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{x}_1^\top - \boldsymbol{x}^\top \\ \vdots \\ \boldsymbol{x}_n^\top - \boldsymbol{x}^\top \end{pmatrix} = \begin{pmatrix} x_{11} - x_1 & \cdots & x_{1p} - x_p \\ \vdots & & \vdots \\ x_{n1} - x_1 & \cdots & x_{np} - x_p \end{pmatrix}$$

で定義する. 従って, 求めるべき $\boldsymbol{a}$ は行列 $\boldsymbol{X}^\top\boldsymbol{X}$ の最大固有値に対する固有ベクトルで長さ 1 のものであり, このとき $f(\boldsymbol{a})$ はその最大固有値に一致する.

このようにして求まったベクトル $\boldsymbol{a}$ を**第 1 主成分方向** *(first principal component direction)* または**第 1 主成分負荷量** *(first principal component loading)* と呼ぶ. また, 得られた特徴量

$$z_{i1} = a_1 x_{i1} + \cdots + a_p x_{ip} \qquad (i = 1, \ldots, n)$$

を**第 1 主成分 (得点)***(first principal component (score))* と呼ぶ.

6.2.2. $d \geq 2$ の場合. まず記号を準備する. $\boldsymbol{X}^\top\boldsymbol{X}$ は非負定値対称行列だから, その固有値はすべて 0 以上の実数である. そこで, $\boldsymbol{X}^\top\boldsymbol{X}$ の固有値を (重複を許して) 降順に並べて $\lambda_1 \geq \cdots \geq \lambda_p (\geq 0)$ と書くことにする. さらに, 各 $j = 1, \ldots, p$ について $\boldsymbol{a}_j$ を λ_j に対する固有ベクトルとする. このとき, $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_p$ はそれぞれ長さ 1 かつ互いに直交するようにとることができる. すなわち,

$$\|\boldsymbol{a}_j\| = 1 \ (j = 1, \ldots, p), \qquad j \neq k \Rightarrow \boldsymbol{a}_j \cdot \boldsymbol{a}_k = 0 \tag{6.1}$$

が成り立つと仮定してよい.[1]

1 つめの特徴量は前節で構成した第 1 主成分を用いる. 前節の議論より, ベクトル $\boldsymbol{a}_1$ は第 1 主成分方向に対応する. 第 1 主成分方向に関してデータが有する情報はベ

[1] 「対称行列は直交行列によって対角化できる (参考文献 1. 系 5.4.6)」という事実の言い換えである. 参考文献 4. の IV 章定理 4' を参照すること.

クトル $(\boldsymbol{a}_1 \cdot \boldsymbol{x}_i)\boldsymbol{a}_1$ $(i = 1, \ldots, n)$ にすべて縮約されているので, 第 1 主成分 $\boldsymbol{a}_1 \cdot \boldsymbol{x}_i$ $(i = 1, \ldots, n)$ にすべて含まれている. 従って, 2 つめの特徴量を構成する指針として, 観測データから第 1 主成分方向の成分を取り除いたデータ

$$\tilde{\boldsymbol{x}}_i := \boldsymbol{x}_i - (\boldsymbol{a}_1 \cdot \boldsymbol{x}_i)\boldsymbol{a}_1 \qquad (i = 1, \ldots, n)$$

に対して, 前節と同様の考え方でこれらのデータたちのばらつきが最も大きい方向 $\boldsymbol{a}$ を求めて, 特徴量 $\boldsymbol{a} \cdot \boldsymbol{x}_i$ $(i = 1, \ldots, n)$ を構成するのが自然である. すなわち,

$$\sum_{i=1}^{n} (\boldsymbol{a} \cdot \tilde{\boldsymbol{x}}_i - \boldsymbol{a} \cdot \bar{\tilde{\boldsymbol{x}}})^2 \qquad \text{ただし} \quad \bar{\tilde{\boldsymbol{x}}} := \frac{1}{n} \sum_{i=1}^{n} \tilde{\boldsymbol{x}}_i$$

を制約条件 $\|\boldsymbol{a}\| = 1$ の下で最大化するような $\boldsymbol{a}$ を選べばよい. 前節と同様に, 行列 $\boldsymbol{X}_{(-1)}$ を

$$\boldsymbol{X}_{(-1)} = \begin{pmatrix} \tilde{\boldsymbol{x}}_1^\top - \bar{\boldsymbol{x}}^\top - (\boldsymbol{a}_1 \cdot \bar{\boldsymbol{x}})\boldsymbol{a}_1^\top \\ \vdots \\ \tilde{\boldsymbol{x}}_n^\top - \bar{\boldsymbol{x}}^\top - (\boldsymbol{a}_1 \cdot \bar{\boldsymbol{x}})\boldsymbol{a}_1^\top \end{pmatrix}$$

で定義すれば, $\boldsymbol{a}$ は行列 $\boldsymbol{X}_{(-1)}^\top \boldsymbol{X}_{(-1)}$ の最大固有値に対する固有ベクトルとなることがわかる. ここで, 行列 $\boldsymbol{X}_{(-1)}$ は以下のように書けることに注意する (E_p は p 次単位行列):

$$\boldsymbol{X}_{(-1)} = \boldsymbol{X} - \boldsymbol{X}\boldsymbol{a}_1\boldsymbol{a}_1^\top = \boldsymbol{X}(E_p - \boldsymbol{a}_1\boldsymbol{a}_1^\top).$$

従って,

$$\boldsymbol{X}_{(-1)}^\top \boldsymbol{X}_{(-1)} = (E_p - \boldsymbol{a}_1\boldsymbol{a}_1^\top)\boldsymbol{X}^\top \boldsymbol{X}(E_p - \boldsymbol{a}_1\boldsymbol{a}_1^\top)$$

であるから, 条件 (6.1) より

$$\boldsymbol{X}_{(-1)}^\top \boldsymbol{X}_{(-1)}\boldsymbol{a}_1 = \boldsymbol{0}, \qquad \boldsymbol{X}_{(-1)}^\top \boldsymbol{X}_{(-1)}\boldsymbol{a}_j = \lambda_j \boldsymbol{a}_j \ (j = 2, \ldots, p)$$

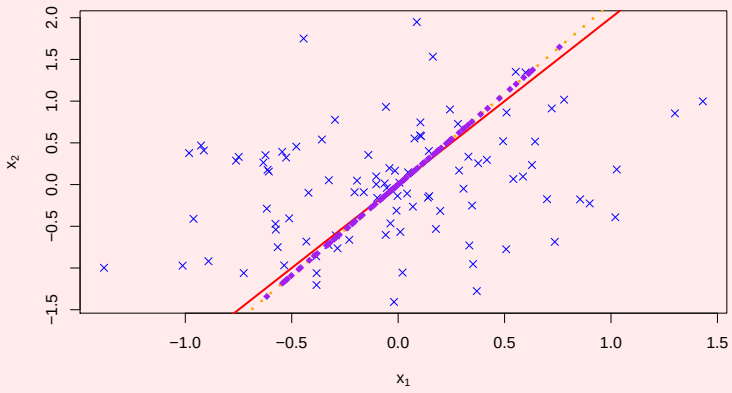
が成り立つ. これは $0, \lambda_2, \ldots, \lambda_p$ が $\boldsymbol{X}_{(-1)}^\top \boldsymbol{X}_{(-1)}$ の固有値であり, $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_p$ がそれぞれに対する固有ベクトルであることを意味する. 従って, $\boldsymbol{X}_{(-1)}^\top \boldsymbol{X}_{(-1)}$ の最大固有値は λ_2 であり, 求めるべきベクトルは $\boldsymbol{a}_2$ である. 前節と同様に, $\boldsymbol{a}_2$ を**第 2 主成分方向**または**第 2 主成分負荷量**と呼び, 得られた特徴量 $\boldsymbol{a}_2 \cdot \boldsymbol{x}_i$ $(i = 1, \ldots, n)$ を**第 2 主成分 (得点)** と呼ぶ.

同様の議論を繰り返すことによって, k 番目の特徴量 (**第 k 主成分 (得点)**) として $\boldsymbol{a}_k \cdot \boldsymbol{x}_i$ $(i = 1, \ldots, n)$ を用いることが自然である. $\boldsymbol{a}_k$ は**第 k 主成分方向**または**第 k 主成分負荷量**と呼ばれる.

6.2.3. R での実行. R には主成分分析を実行するための関数として, 関数 `prcomp()` および関数 `princomp()` が用意されている. 前者と後者には計算法に若干の違いがあり, 一般には前者の方が数値計算の観点からみると優れている (後者は S 言語との互換性を重視した実装となっている). 従って以下では関数 `prcomp()` を利用する.

```
> # 人工データによる確認 (2 次元)
> set.seed(123)
> n <- 100 # サンプル数
> (a <- c(1, 2)/sqrt(5)) # 主成分方向
[1] 0.4472136 0.8944272
> x <- runif(n, -1, 1) %o% a + rnorm(2*n, sd = 0.5) # 観測データ
> ## a のスカラー倍に正規乱数がのった形となっており,
> ## a 方向に本質的な情報が集約されていることがわかる
> plot(x, pch = 4, col = "blue", xlab = expression(x[1]),
+      ylab = expression(x[2])) # 散布図
> abline(0, a[2]/a[1], col = "red", lwd = 2)
```

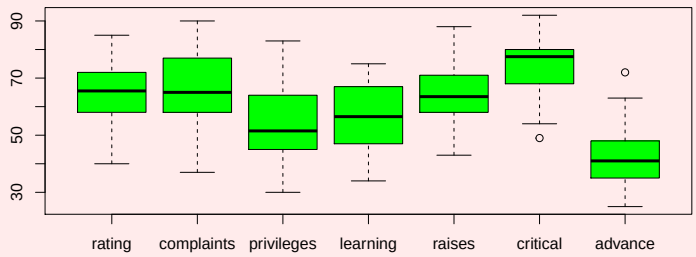
```
> ## 主成分方向に対応する直線を上書き
> (res <- prcomp(x)) # 主成分分析
Standard deviations (1, .., p=2):
[1] 0.7128168 0.4951924

Rotation (n x k) = (2 x 2):
           PC1        PC2
[1,] -0.4178880 -0.9084985
[2,] -0.9084985  0.4178880
> ## 第 1 主成分方向が a に非常に近い (符号は反対)
> abline(0, res$rotation[2,1]/res$rotation[1,1], col = "orange",
+        lty = "dotted", lwd = 3)
> ## 推定された第 1 主成分方向に対応する直線を上書き
> pc1 <- predict(res)[ ,1] # 第 1 主成分
> points(pc1 %o% res$rotation[ ,1], pch = 18, col = "purple") # 第 1 主成分を上書き
> # 固有値分解との比較
> (out <- eigen(crossprod(scale(x,scale = FALSE)))) # 固有値分解
eigen() decomposition
$values
[1] 50.30267 24.27634

$vectors
          [,1]       [,2]
[1,] 0.4178880 -0.9084985
[2,] 0.9084985  0.4178880
> out$vectors # 符号を除いて主成分負荷量と一致
          [,1]       [,2]
[1,] 0.4178880 -0.9084985
[2,] 0.9084985  0.4178880
> sqrt(out$values/(n-1)) # 主成分の標準偏差に対応
[1] 0.7128168 0.4951924
> res$sdev
[1] 0.7128168 0.4951924
```



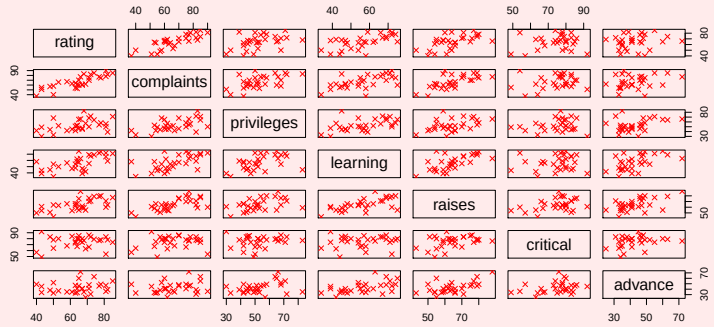
(pca-simulate.r)

```
> ## データセット attitude による例
> ## ある金融機関の 30 の部署の事務職員への管理者の態度に関する
> ## アンケート調査．以下の 7 つの質問からなり，好意的な回答を
> ## したパーセントからなる
```

```
> ## rating: 全体的な評価
> ## complaints: 職員の苦情への対処
> ## privileges: えこひいきしていないか
> ## learning: 学習の機会はあるか
> ## raises: 仕事にみあった昇給をしているか
> ## critical: 批判的すぎないか
> ## advance: 昇進の機会はあるか
> boxplot(attitude, col = "green") # 箱ひげ図
```




```
> plot(attitude, pch = 4, col = "red") # 散布図
```



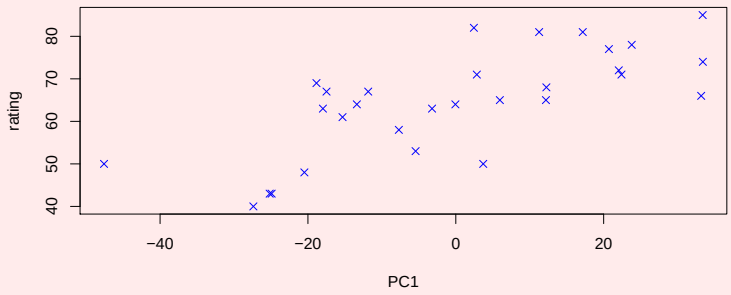

```
> dat <- subset(attitude, select = -rating)
> # rating は全体の特徴量に相当するので分析から除外
> res <- prcomp(dat) # 主成分分析
> res$rotation # 各主成分方向を列ベクトルとする行列
                PC1        PC2        PC3        PC4        PC5         PC6
complaints 0.5416077  0.43545089 -0.3992124  0.3334014 -0.2124872  0.44873797
privileges 0.4385746  0.32538210  0.1034507 -0.8197242  0.0134477 -0.13765240
learning   0.4681789 -0.03909117  0.4250002  0.3072272  0.7088962 -0.04162848
raises     0.4249311 -0.23035630 -0.1128799  0.2407430 -0.3737876 -0.74562279
critical   0.1591573 -0.55995951 -0.6746707 -0.2332684  0.3755895  0.10223589
advance    0.2987048 -0.57996742  0.4258872 -0.1006769 -0.4139444  0.45994304
> ### 第 1 主成分方向は全変数の符号が同じであり，全体の傾向を表す方向と解釈できる
> ### 第 2 主成分方向は正の向きに管理者の能力を表す変数，負の向きに職員の待遇を表す変数がきている
> ### 第 3 主成分方向は正の向きにポジティブな項目，負の向きにネガティブな項目を表す変数がきている
> ### と解釈できそう
> z <- predict(res) # 主成分の計算
> head(z)
        PC1        PC2        PC3         PC4       PC5       PC6
1 -24.909261 -23.652930 -13.882809  3.32362046 -2.332185  2.328505
2  -3.212328  -2.726829   2.919654 -0.23572835 -2.890338  2.133129
3  22.407622  -6.023068  -1.154227 -7.56573062  6.306272 -6.043048
```

```
4 -15.320742  -1.967673 -11.792944 -1.32589237  4.742095  5.117589
5  17.172220  -2.915145  -4.701811  3.61325655  3.474427  2.285034
6 -25.173052  16.146072  13.727144  0.07139001 -8.362592 -2.936238
```

> *colMeans(z) # prcomp* はデータの平均を差し引いてから計算するため，主成分の平均は *0*

```
          PC1           PC2           PC3           PC4           PC5
 1.328567e-15  1.702342e-15 -1.058413e-15 -9.631185e-16  0.000000e+00
          PC6
 9.144537e-15
```

> *plot(attitude$rating ~ z[,1], xlab = "PC1", ylab = "rating",*
> *+ pch = 4, col = "blue") # rating* と第 *1* 主成分の関係



> *prcomp(~. -rating, data = attitude) #* 上と同じ分析を実行

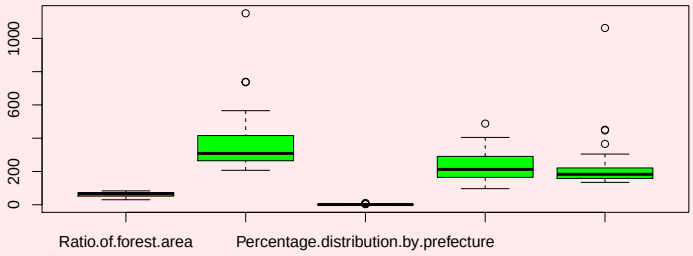
```
Standard deviations (1, .., p=6):
[1] 20.699356 10.727851  9.465937  8.938540  6.292552  4.895626

Rotation (n x k) = (6 x 6):
                PC1         PC2        PC3        PC4        PC5         PC6
complaints 0.5416077  0.43545089 -0.3992124  0.3334014 -0.2124872  0.44873797
privileges 0.4385746  0.32538210  0.1034507 -0.8197242  0.0134477 -0.13765240
learning   0.4681789 -0.03909117  0.4250002  0.3072272  0.7088962 -0.04162848
raises     0.4249311 -0.23035630 -0.1128799  0.2407430 -0.3737876 -0.74562279
critical   0.1591573 -0.55995951 -0.6746707 -0.2332684  0.3755895  0.10223589
advance    0.2987048 -0.57996742  0.4258872 -0.1006769 -0.4139444  0.45994304
```

> *prcomp(dat, scale. = TRUE) #* 各変数をその標準偏差で割って分析を実行（結果は変わる）

```
Standard deviations (1, .., p=6):
[1] 1.7802312 1.0031684 0.8734465 0.7433145 0.5632464 0.4379022

Rotation (n x k) = (6 x 6):
                PC1        PC2          PC3         PC4         PC5
complaints 0.4393752 -0.3126424  0.445166951 -0.31601946  0.19152122
privileges 0.3947108 -0.3087507  0.217413750  0.81484689  0.03768625
learning   0.4614010 -0.2170870 -0.271981397 -0.22479562 -0.77564752
raises     0.4926576  0.1155323  0.005604908 -0.36510795  0.46036381
critical   0.2248130  0.8022474  0.457245609  0.09994698 -0.28887525
advance    0.3808011  0.3207060 -0.686643142  0.20574245  0.25472836
                 PC6
complaints  0.61194923
privileges -0.19029420
learning   -0.11767060
raises     -0.63140375
critical    0.05784728
advance     0.41646475
```

> *### 各変数のばらつきに大きな違いがある場合，ばらつきが大きい変数の*
> *### 効果を重視してしまうことになるため，全変数を同等に扱うためには*
> *### scale. = TRUE* として分析を実行すべきである

```
> ## 総務省統計局の統計データ
> ## http://www.stat.go.jp/data/shihyou/naiyou.htm
> ## 社会生活統計指標－都道府県の指標－ 2017 社会生活統計指標 2017 年 2 月 17 日公表
> ## http://www.e-stat.go.jp/SG1/estat/List.do?bid=000001083999&cycode=0
> ## 森林面積割合 Ratio of forest area (%) 2014
> ## 就業者 1 人当たり農業産出額 (販売農家)
> ## Gross agricultural product per agricultural worker (commercial farm households)
> ## (万円:10 thousand yen) 2014
> ## 全国総人口に占める人口割合
> ## Percentage distribution by prefecture (%) 2015
> ## 土地生産性（耕地面積 1 ヘクタール当たり）
> ## Land productivity (per hectare of cultivated land area)
> ## (万円:10 thousand yen) 2014
> ## 商業年間商品販売額［卸売業＋小売業］(事業所当たり)
> ## Annual sales of commercial goods        [wholesale and retail trade] (per establishment)
> ## (百万円:million yen) 2013
> ### データの読み込み
> kendata <- read.csv(file="kendata.csv",row.names=1,header = TRUE)
> ### 注意：csv ファイルは作業ディレクトリ (コンソールの上部に書いてある)
> ### においてある必要がある
> head(kendata)

         Ratio.of.forest.area Gross.agricultural.product
Hokkaido                 67.9                      1150.6
Aomori                   63.8                       444.7
Iwate                    74.9                       334.3
Miyagi                   55.9                       299.9
Akita                    70.5                       268.7
Yamagata                 68.7                       396.3
         Percentage.distribution.by.prefecture Land.productivity
Hokkaido                                  4.23              96.8
Aomori                                    1.03             186.0
Iwate                                     1.01             155.2
Miyagi                                    1.84             125.3
Akita                                     0.81              98.5
Yamagata                                  0.88             174.1
         Annual.sales.of.commercial.goods
Hokkaido                            283.3
Aomori                              183.0
Iwate                               179.4
Miyagi                              365.9
Akita                               153.3
Yamagata                            157.5

> boxplot(kendata, col = "green") # 変数のばらつきに大きな違いがある
```




```
> (mypca <- prcomp(kendata,scale=TRUE))
```

```
Standard deviations (1, .., p=5):
[1] 1.5903931 1.0698965 0.8195653 0.7076020 0.3918975

Rotation (n x k) = (5 x 5):
                                          PC1        PC2         PC3
Ratio.of.forest.area                -0.4871498  0.1045813 -0.45748795
Gross.agricultural.product           0.1339190  0.8115056  0.47912767
Percentage.distribution.by.prefecture  0.5851294 -0.1511042  0.04467249
Land.productivity                    0.3547649  0.4851374 -0.74167904
Annual.sales.of.commercial.goods     0.5258481 -0.2689436 -0.09517368
                                          PC4         PC5
Ratio.of.forest.area                 0.6859649 -0.26815060
Gross.agricultural.product           0.3045447  0.03483694
Percentage.distribution.by.prefecture  0.1640953 -0.77837539
Land.productivity                   -0.2897485  0.06885892
Annual.sales.of.commercial.goods     0.5708093  0.56238052
> ### 第 1 主成分は都会度を表す成分 (正の向きほど高い)
> ### 第 2 主成分は産業の中心が農業か商業か (正の向きほど高い)
> ### と解釈できそう
```

(pca.r)

6.3. 分析の評価

6.3.1. 寄与率. 構成した主成分が元のデータがもっていた情報をどの程度保持しているかを評価する方法の 1 つとして, 回帰分析の場合と同様に, その主成分のばらつき (分散) がもとのデータのばらつき (分散) をどの程度説明できているか評価する方法がある. 第 k 主成分に対しては, これは

$$\frac{\frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{a}_k\cdot\boldsymbol{x}_i-\boldsymbol{a}_k\cdot\bar{\boldsymbol{x}})^2}{\frac{1}{n-1}\sum_{i=1}^{n}\|\boldsymbol{x}_i-\bar{\boldsymbol{x}}\|^2}$$

を計算することに相当する. この量を第 k 主成分の**寄与率** *(proportion of variance)* と呼ぶ. $\boldsymbol{a}_k\cdot\boldsymbol{x}_i-\boldsymbol{a}_k\cdot\bar{\boldsymbol{x}}$ が (列) ベクトル $\boldsymbol{X}\boldsymbol{a}$ の第 i 成分に対応することと, 列ベクトル $\boldsymbol{v}$ に対して $\|\boldsymbol{v}\|^2=\boldsymbol{v}^\top\boldsymbol{v}$ が成り立つことに注意すれば,

$$\sum_{i=1}^{n}(\boldsymbol{a}_k\cdot\boldsymbol{x}_i-\boldsymbol{a}_k\cdot\bar{\boldsymbol{x}})^2=\|\boldsymbol{X}\boldsymbol{a}_k\|^2=\boldsymbol{a}_k^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{a}_k=\lambda_k\boldsymbol{a}_k^\top\boldsymbol{a}_k=\lambda_k\|\boldsymbol{a}_k\|^2=\lambda_k$$

を得る. さらに, $\|\boldsymbol{x}_i-\bar{\boldsymbol{x}}\|^2$ が行列 $\boldsymbol{X}\boldsymbol{X}^\top$ の第 i 対角成分であることと, 直交行列 A を $A=(\boldsymbol{a}_1,\ldots,\boldsymbol{a}_p)$ で定義すれば,

$$A^\top\boldsymbol{X}\boldsymbol{X}A=\begin{pmatrix}\lambda_1 & & \\ & \ddots & \\ & & \lambda_p\end{pmatrix}$$

と対角化されることに注意すれば,

$$\sum_{i=1}^{n}\|\boldsymbol{x}_i-\bar{\boldsymbol{x}}\|^2=\mathrm{tr}(\boldsymbol{X}\boldsymbol{X}^\top)=\mathrm{tr}(\boldsymbol{X}^\top\boldsymbol{X})=\mathrm{tr}(\boldsymbol{X}^\top\boldsymbol{X}E_p)=\mathrm{tr}(\boldsymbol{X}^\top\boldsymbol{X}AA^\top)$$

$$=\mathrm{tr}(A^\top\boldsymbol{X}^\top\boldsymbol{X}A)=\sum_{l=1}^{p}\lambda_l$$

が成り立つ. ここで, 積 BC および CB が定義されるような行列 B,C に対して $\mathrm{tr}(BC)=\mathrm{tr}(CB)$ が成り立つことを用いた. 以上より, 第 k 主成分の寄与率は,

$$\frac{\lambda_k}{\sum_{l=1}^{p}\lambda_l}$$

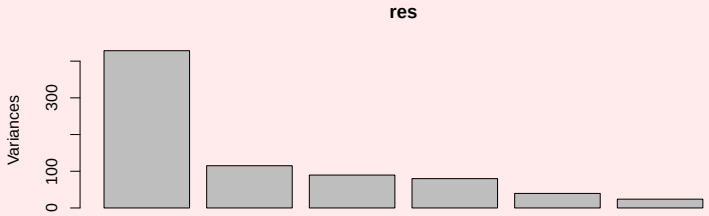
で計算される. 第 1 主成分から第 k 主成分までを特徴量として用いた際に説明できるデータのばらつきの割合は第 k 主成分までの**累積寄与率** *(cumulative proportion)* と呼ばれ, 第 1 主成分の寄与率から第 k 主成分の寄与率までの総和として計算される:

$$\frac{\sum_{l=1}^{k} \lambda_l}{\sum_{l=1}^{p} \lambda_l}.$$

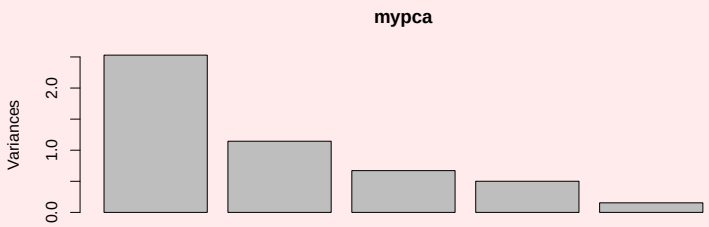
累積寄与率はいくつの主成分を用いるべきかの基準として用いられる. 一般に, 累積寄与率が 80%程度の主成分を使って分析を行うことが多い.

R では, 寄与率および累積寄与率は, 関数 `prcomp()` のアウトプットに関数 `summary()` を適用することで計算できる.

```
> ## データセット attitude による例
> res <- prcomp( ~. -rating, data = attitude)
> summary(res) # 第 3 主成分までの累積寄与率が 80%を超えている
Importance of components:
                          PC1     PC2    PC3    PC4     PC5     PC6
Standard deviation     20.6994 10.7279 9.4659 8.9385 6.29255 4.89563
Proportion of Variance  0.5517  0.1482 0.1154 0.1029 0.05099 0.03086
Cumulative Proportion   0.5517  0.6999 0.8153 0.9182 0.96914 1.00000
> plot(res) # スクリープロット (各主成分の分散を棒グラフで表示)
```




```
> ## kendata.csv による例
> ### データの読み込み
> kendata <- read.csv(file="kendata.csv",row.names=1,header = TRUE)
> mypca <- prcomp(kendata, scale=TRUE)
> summary(mypca) # 第 3 主成分までの累積寄与率が 80%を超えている
Importance of components:
                          PC1    PC2    PC3    PC4     PC5
Standard deviation     1.5904 1.0699 0.8196 0.7076 0.39190
Proportion of Variance 0.5059 0.2289 0.1343 0.1001 0.03072
Cumulative Proportion  0.5059 0.7348 0.8691 0.9693 1.00000
> plot(mypca) # スクリープロット
```

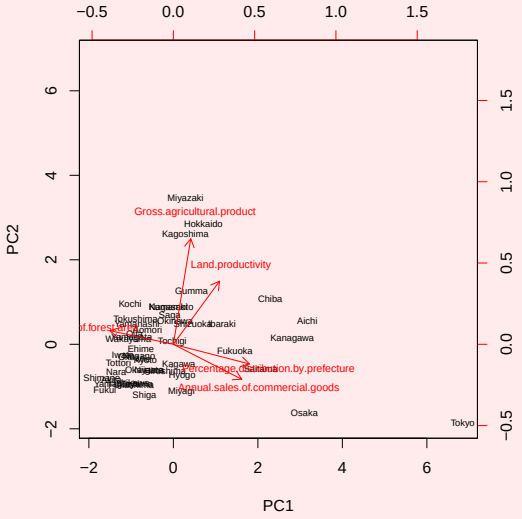


(pca-summary.r)

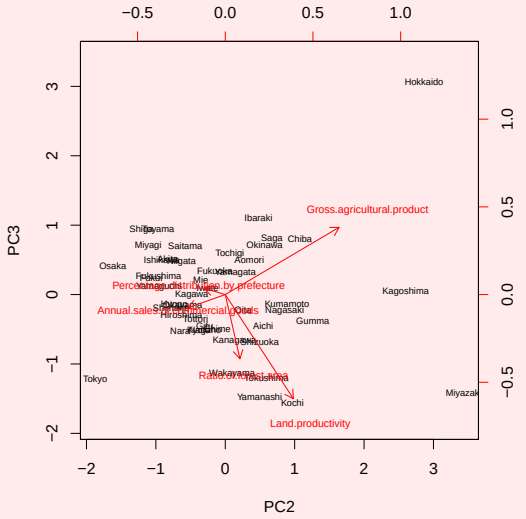
6.3.2. バイプロット. 関連がある 2 枚の散布図を 1 つの画面に表示する散布図を**バイプロット** *(biplot)* という. 主成分分析では, 得られた主成分の意味を解釈するために, 主成分方向の散布図と主成分の散布図を対応づけて分析を進める場合が多い. より具体的には, 2 つの主成分方向 $\boldsymbol{a}_k = (a_{1k}, \ldots, a_{pk})^\top$ と $\boldsymbol{a}_l = (a_{1l}, \ldots, a_{pl})^\top$ に着目する. 主成分方向の散布図とは点 $\{(a_{jk}, a_{jl})\}_{j=1}^p$ の散布図であり, 各変数が着目した主成分方向にどれだけの変動成分をもつかを図示している (主成分分析ではこれらの点を対応するベクトルで描画することが多い). 一方で, 主成分の散布図とは本質的に点 $\{(\boldsymbol{a}_k \cdot (\boldsymbol{x}_i - \bar{\boldsymbol{x}}), \boldsymbol{a}_l \cdot (\boldsymbol{x}_i - \bar{\boldsymbol{x}})\}_{i=1}^n$ の散布図であり, 各サンプルが着目した主成分方向にどれだけの変動成分をもつかを図示している.

R では, 関数 `prcomp()` のアウトプットに関数 `biplot()` を適用することでバイプロットを描画できる.

```
> ## kendata.csv による例
> ### データの読み込み
> kendata <- read.csv(file="kendata.csv",row.names=1,header = TRUE)
> mypca <- prcomp(kendata, scale=TRUE) # 主成分分析の実行
> biplot(mypca,cex=c(0.6, 0.7), scale = 0) # バイプロット (第 1 vs 第 2 主成分)
> ### 第 1 主成分方向の正の向きには大都市をもつ県が集中
> ### 人口割合, 商品販売額および森林面積割合は 1 人当たり農業産出額とほぼ直交しており,
> ### 両者に関連はあまりないといえそう
> ### 第 2 主成分方向の正の向きには 1 人当たり農業産出額の上位県が集中
> gross <- setNames(kendata$Gross.agricultural.product, row.names(kendata))
> head(sort(gross,decreasing=TRUE))
 Hokkaido  Miyazaki Kagoshima     Chiba     Gumma   Ibaraki
   1150.6     739.1     736.5     565.5     530.6     479.0
```

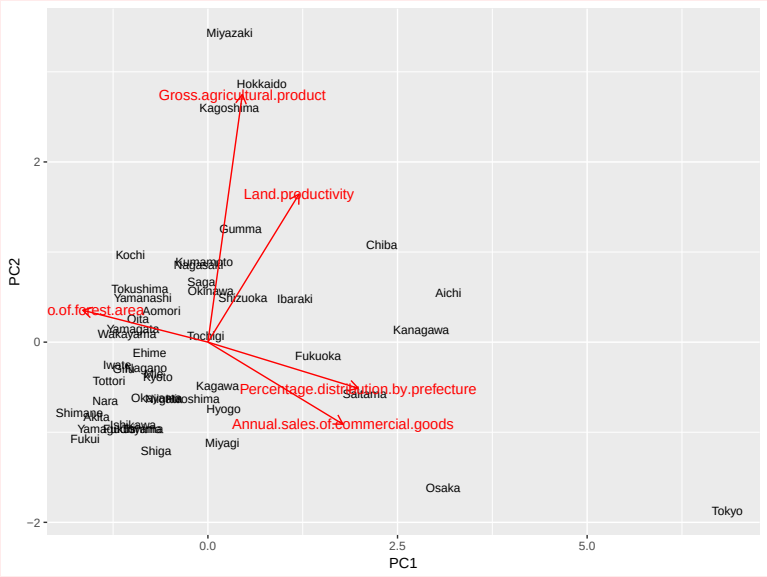



```
> biplot(mypca,choices=c(2,3),cex=c(0.6, 0.7), scale = 0) # バイプロット (第 2 vs 第 3 主成分)
> ### 第 3 主成分方向の負の向きには土地生産性の上位県が集中
> land <- setNames(kendata$Land.productivity, row.names(kendata))
> head(sort(land,decreasing=TRUE))
 Miyazaki     Tokyo  Kanagawa     Aichi Kagoshima     Kochi
    487.7     404.7     396.4     388.9     351.2     339.9
> ### 北海道の土地生産性は低い
> head(sort(land))
Hokkaido    Akita   Toyama    Fukui    Shiga Ishikawa
    96.8     98.5     98.5     98.5    104.9    112.0
```




```
> ## パッケージ ggfortify によるバイプロット
> install.packages("ggfortify") # パッケージのインストール
```

```
> library(ggfortify) # パッケージのロード
> autoplot(mypca, shape=FALSE, label=TRUE, loadings=TRUE,
+          loadings.label=TRUE,label.size=3, scale = 0,
+          loadings.label.size=4) # バイプロット（第 1 vs 第 2 主成分）
```




```
> library(ggfortify) # パッケージのロード
> autoplot(mypca, shape=FALSE, label=TRUE, loadings=TRUE,
+          loadings.label=TRUE,label.size=3, scale = 0,
+          loadings.label.size=4, x = 2, y = 3) # バイプロット（第 2 vs 第 3 主成分）
```

(pca-biplot2.r)

6.4. 参考文献

1. 二木昭人著「基礎講義 線形代数学」, 培風館 (1999 年).
2. G. James, D. Witten, T. Hastie, R. Tibshirani 著「An Introduction to Statistical Learning」, Springer (2013 年).
3. 金明哲著「R によるデータサイエンス (第 2 版)」, 森北出版 (2017 年).
4. 佐竹一郎著「線型代数学」, 裳華房 (1973 年).
5. 杉浦光夫著「解析入門 I」, 東京大学出版会 (1980 年).
6. 杉浦光夫著「解析入門 II」, 東京大学出版会 (1985 年).

6.5. 補足: 主成分分析の計算法の詳細 ($d = 1$ の場合)

$f(\boldsymbol{a})$ は連続関数であり, また集合 $\{\boldsymbol{a} \in \mathbb{R}^p : \|\boldsymbol{a}\| = 1\}$ はコンパクト (有界閉集合) であるから, この最大化問題は解をもつ.[2] 更に, Lagrange の乗数法から, 求めるべき解は, Lagrange 関数

$$L(\boldsymbol{a}, \lambda) = f(\boldsymbol{a}) + \lambda(1 - \|\boldsymbol{a}\|^2)$$

の勾配を 0 にするベクトルである.[3] いま, $\bar{\boldsymbol{x}} = (\bar{x}_1, \ldots, \bar{x}_p)^\top$, すなわち

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} \qquad (j = 1, \ldots, p)$$

と書くことにすると,

$$f(\boldsymbol{a}) = \sum_{i=1}^{n} \left(\sum_{j=1}^{p} a_j (x_{ij} - \bar{x}_j) \right)^2$$

[2] 参考文献 5., I 章定理 7.3 参照.
[3] 参考文献 6., VI 章定理 3.1 系参照.

より, 各 $j = 1, \ldots, p$ について

$$\frac{\partial L}{\partial a_j}(\boldsymbol{a}, \lambda) = 2\sum_{i=1}^{n}\left(\sum_{k=1}^{p} a_k(x_{ik} - \bar{x}_k)\right)(x_{ij} - \bar{x}_j) - 2\lambda a_j$$

$$= 2\sum_{k=1}^{p}\left(\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)\right)a_k - 2\lambda a_j \tag{6.2}$$

が成り立つ. (6.2) の右辺第 1 項は p 次元ベクトル $\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{a}$ の第 j 成分に等しいことがわかる. 以上より, 求めるべき $\boldsymbol{a}$ は, 方程式

$$\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{a} = \lambda \boldsymbol{a} \tag{6.3}$$

の解となることがわかる. 特に, λ は p 次正方行列 $\boldsymbol{X}^\top \boldsymbol{X}$ の固有値であり, $\boldsymbol{a}$ は λ に対する固有ベクトルとなる. また, (6.3) の両辺に左から $\boldsymbol{a}^\top$ をかけると,

$$\boldsymbol{a}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{a} = \lambda \boldsymbol{a}^\top \boldsymbol{a} = \lambda \|\boldsymbol{a}\|^2 = \lambda$$

を得る. ここで,

$$\boldsymbol{X}\boldsymbol{a} = ((\boldsymbol{x}_1 - \bar{\boldsymbol{x}}) \cdot \boldsymbol{a}, \ldots, (\boldsymbol{x}_n - \bar{\boldsymbol{x}}) \cdot \boldsymbol{a})^\top$$

であることに注意すれば,

$$\boldsymbol{a}^\top \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{a} = (\boldsymbol{X}\boldsymbol{a})^\top \boldsymbol{X}\boldsymbol{a} = \|\boldsymbol{X}\boldsymbol{a}\|^2 = f(\boldsymbol{a})$$

を得る. このことから, $f(\boldsymbol{a})$ は行列 $\boldsymbol{X}^\top \boldsymbol{X}$ の固有値となる. 以上をまとめると, 制約条件 $\|\boldsymbol{a}\| = 1$ の下で関数 $f(\boldsymbol{a})$ を最大化するようなベクトル $\boldsymbol{a}$ は存在して次の性質をもつ:

- $f(\boldsymbol{a})$ は行列 $\boldsymbol{X}^\top \boldsymbol{X}$ の固有値であり, $\boldsymbol{a}$ はこの固有値に対する固有ベクトルである.