応用統計学 2017 主成分分析に関する補足資料

2017 年 10 月 18 日 (水)

清 智也 sei@mist.i.u-tokyo.ac.jp

http://ur0.pw/yTzt

先週の講義で主成分分析について少し混乱したので，整理しておく。

まず，データ行列を

$$\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p) = \begin{pmatrix} \boldsymbol{x}^{(1)\prime} \\ \vdots \\ \boldsymbol{x}^{(n)\prime} \end{pmatrix} \in \mathbb{R}^{n \times p}$$

と記す。ただし各列はあらかじめ標準化されているものとする：

$$\bar{x}_i = \frac{1}{n}\mathbf{1}_n'\boldsymbol{x}_i = 0, \quad \mathrm{V}(\boldsymbol{x}_i) = \frac{1}{n}(\boldsymbol{x}_i - \bar{x}_i\mathbf{1}_n)'(\boldsymbol{x}_i - \bar{x}_i\mathbf{1}_n) = 1, \quad i = 1, \ldots, p.$$

先週の講義では，

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}' = \sum_{k=1}^{p} d_k \boldsymbol{u}_k \boldsymbol{v}_k'$$

と特異値分解し，右辺の第 k 項 $d_k\boldsymbol{u}_k\boldsymbol{v}_k'$ を第 k 主成分と呼んだ。しかし一般には変数 $\boldsymbol{x} \in \mathbb{R}^p$ に対して $\boldsymbol{v}_k'\boldsymbol{x}$ のことを $\boldsymbol{x}$ の第 k 主成分と呼ぶことが多いようである[1]。この用語の下で主成分に $\boldsymbol{x} = \boldsymbol{x}^{(t)}$ を代入した値を**主成分得点**と呼ぶ。第 k 主成分の主成分得点は

$$\boldsymbol{v}_k'\boldsymbol{x}^{(t)} = (\boldsymbol{X}\boldsymbol{V})_{tk} = (\boldsymbol{U}\boldsymbol{D})_{tk} = d_k(\boldsymbol{u}_k)_t, \quad t = 1, \ldots, n,$$

となる。これをすべての個体について並べてできるベクトルは $d_k\boldsymbol{u}_k$ となる。

また，主成分得点と各変量の相関係数を**因子負荷量**[2]と呼ぶ。第 k 主成分の因子負荷量は

$$\begin{aligned} \mathrm{Corr}(d_k\boldsymbol{u}_k, \boldsymbol{x}_i) &= \frac{\mathrm{Cov}(d_k\boldsymbol{u}_k, \boldsymbol{x}_i)}{\sqrt{\mathrm{V}(d_k\boldsymbol{u}_k)\mathrm{V}(\boldsymbol{x}_i)}} \\ &= \frac{1}{\sqrt{n}}\boldsymbol{u}_k'\boldsymbol{x}_i \quad (\because\ \bar{u}_k = \bar{x}_i = 0,\ \boldsymbol{u}_k'\boldsymbol{u}_k = 1,\ \mathrm{V}(\boldsymbol{x}_i) = 1) \\ &= \frac{1}{\sqrt{n}} d_k(\boldsymbol{v}_k)_i \end{aligned}$$

となる。これをすべての変量について並べてできるベクトルは

$$\frac{d_k}{\sqrt{n}}\boldsymbol{v}_k$$

となる。

[1] たとえば永田 靖，棟近 雅彦「多変量解析法入門」，サイエンス社。

[2] 主成分負荷量ともいう。

さて，主成分分析の説明では，共分散行列 $\boldsymbol{S} = (1/n)\boldsymbol{X}'\boldsymbol{X}$ を出発点とすることが多い。すなわち，$\boldsymbol{S}$ の第 k 固有ベクトルを $\boldsymbol{v}_k$ とおくとき，ベクトル $\boldsymbol{x} \in \mathbb{R}^p$ の第 k 主成分を $\boldsymbol{v}_k'\boldsymbol{x}$ と定義するのである。この定義は上の定義と一致する。実際，特異値分解を用いると

$$\boldsymbol{S} = \frac{1}{n}\boldsymbol{X}'\boldsymbol{X} = \frac{1}{n}\boldsymbol{V}\boldsymbol{D}\boldsymbol{U}'\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}' = \frac{1}{n}\boldsymbol{V}\boldsymbol{D}^2\boldsymbol{V}'.$$

となり，$\boldsymbol{S}$ のスペクトル分解が得られる。また $\boldsymbol{S}$ の固有値は $\tilde{\lambda}_k := d_k^2/n$ と表され，$\tilde{\lambda}_k$ を用いて因子負荷量を表すと

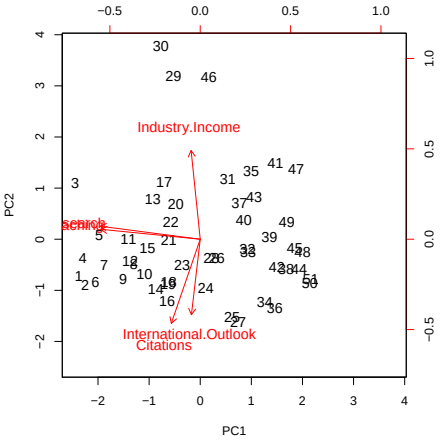
$$\frac{d_k}{\sqrt{n}}\boldsymbol{v}_k = \sqrt{\tilde{\lambda}_k}\boldsymbol{v}_k$$

となる。先週の講義で混乱したのは，この式の左辺と右辺を混同していたことが原因だった。

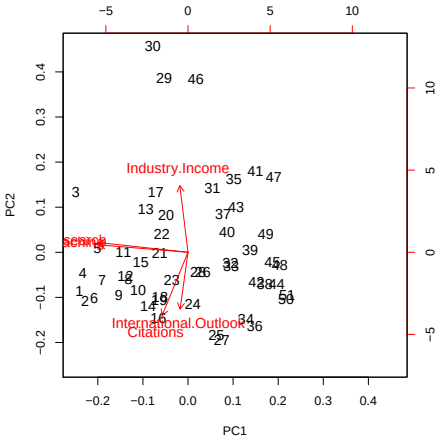
バイプロットに関する補足　先週の講義で説明したバイプロットは，第 1，第 2 主成分得点をプロットした散布図に，変数軸をベクトルとして描き加えたものであった（下図 (a)）。この方法は，幾何学的には個体空間 $\mathbb{R}^p$ において $\boldsymbol{v}_1, \boldsymbol{v}_2$ の張る平面への直交射影を考えている。ただし**個体空間**とは各個体 $\boldsymbol{x}^{(t)}$ が属す空間という意味である[3]。

これに対し，変量空間 $\mathbb{R}^n$ において $\boldsymbol{u}_1, \boldsymbol{u}_2$ の張る平面への直交射影を考えることもできる（下図 (b)）。**変量空間**とは各変量 $\boldsymbol{x}_i$ が属す空間という意味である。

後者の利点は，因子負荷量（の定数倍）を図から読み取れることである。R 言語の関数 `biplot` では，後者がデフォルトで用いられる。また，これら以外の計量を考えることも可能である[4]。どの計量を選んでも，個体と変量の内積 $d_1u_{t1}v_{i1} + d_2u_{t2}v_{i2}$ は不変である。



(a) 個体空間におけるバイプロット
$(d_1\boldsymbol{u}_1, d_2\boldsymbol{u}_2)$ と $(\boldsymbol{v}_1, \boldsymbol{v}_2)$
`biplot(prcomp(X), scale=0)`



(b) 変量空間におけるバイプロット
$(\boldsymbol{u}_1, \boldsymbol{u}_2)$ と $(d_1\boldsymbol{v}_1, d_2\boldsymbol{v}_2)$
`biplot(prcomp(X))`

[3]柴田 里程「データ分析とデータサイエンス」，近代科学社。

[4]K. R. Gabriel, The biplot graphic display of matrices with application to principal component analysis, *Biometrika*, **58** (3), 453–467.