Principal Component Analysis (II) [Sample variation]

Sample principal components

Let $\boldsymbol{X}_j \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma}), j = 1, \ldots, n$ be a random sample and $\boldsymbol{X}_j = \left(X_{1j}, X_{2j}, \ldots, X_{pj}\right)'$.

Also, denote the data by $\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1' \\ \vdots \\ \boldsymbol{X}_p' \end{pmatrix}$

Further, let $\overline{\boldsymbol{X}}$ be the sample mean and $\boldsymbol{S}$ be the sample variance which is an unbiased estimator of $\boldsymbol{\Sigma}$.

Assume that $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Yp)'$ where Y_i is a linear combination of $\boldsymbol{X}_j$. That is $Y_i = \boldsymbol{l}_i' \boldsymbol{X}_j$. Using the same principal component techniques introduced earlier for population principal components and based on the sample variance $\boldsymbol{S}$, can then obtain a set of sample principal components $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Yp)'$.

The data $\boldsymbol{X}$ can be transformed using the principal components. The transformed dataset is

$$\boldsymbol{W} = \boldsymbol{X\Gamma}$$

where $\boldsymbol{\Gamma} = [\boldsymbol{l}_1 \quad \cdots \quad \boldsymbol{l}_p]$. Therefore, the transformed data has sample mean $\overline{\boldsymbol{X}}\boldsymbol{\Gamma}$ and sample variance-covariance $\boldsymbol{\Gamma}'\boldsymbol{S\Gamma}$

Sample principal components (Example: from Johnson and Wichern)

Tract (5) information on 5 socioeconomic variables in Wisconsin. The five variables are

1. Total population (thousands)
2. Median school years
3. Total employment (thousands)
4. Health services employment (hundreds)
5. Median home value ($10,000)

Data:

```
> a <- read.csv("D:\\teaching\\course\\18-19\\STAT5103\\lecture\\PC\\tract.csv",header=T)
> a
   track popul education employment health homevalue
1      1 5.935      14.2      2.265   2.27      2.91
2      2 1.523      13.1      0.597   0.75      2.62
3      3 2.599      12.7      1.237   1.11      1.72
4      4 4.009      15.2      1.649   0.81      3.02
5      5 4.687      14.7      2.312   2.50      2.22
6      6 8.044      15.6      3.641   4.51      2.36
7      7 2.766      13.3      1.244   1.03      1.97
8      8 6.538      17.0      2.618   2.39      1.85
9      9 6.451      12.9      3.147   5.52      2.01
10    10 3.314      12.2      1.606   2.18      1.82
11    11 3.777      13.0      2.119   2.83      1.80
12    12 1.530      13.8      0.798   0.84      4.25
13    13 2.768      13.6      1.336   1.75      2.64
14    14 6.585      14.9      2.763   1.91      3.17
```

Sample principal components (Example: from Johnson and Wichern)

```
> track <- a[,-1]
> track
  popul education employment health homevalue
1    5.935     14.2      2.265   2.27      2.91
2    1.523     13.1      0.597   0.75      2.62
3    2.599     12.7      1.237   1.11      1.72
4    4.009     15.2      1.649   0.81      3.02
5    4.687     14.7      2.312   2.50      2.22
6    8.044     15.6      3.641   4.51      2.36
7    2.766     13.3      1.244   1.03      1.97
8    6.538     17.0      2.618   2.39      1.85
9    6.451     12.9      3.147   5.52      2.01
10   3.314     12.2      1.606   2.18      1.82
11   3.777     13.0      2.119   2.83      1.80
12   1.530     13.8      0.798   0.84      4.25
13   2.768     13.6      1.336   1.75      2.64
14   6.585     14.9      2.763   1.91      3.17

> is.matrix(track)
[1]FALSE    #not a matrix and cannot perform matrix operations on track

> X <- as.matrix(track)
```

Sample principal components (Example: from Johnson and Wichern)

Use sample covariance for principal component analysis

```
> eigen(cov(track))
$values
[1] 6.93107360  1.78514434  0.38964992  0.22952892  0.01415498

$vectors
            [,1]         [,2]          [,3]         [,4]          [,5]
[1,]  0.78120807  -0.07087183   0.003656607   0.54171007   0.302039670
[2,]  0.30564856  -0.76387277  -0.161817438  -0.54479937   0.009279632
[3,]  0.33444840   0.08290788   0.014841008   0.05101636  -0.937255367
[4,]  0.42600795   0.57945799   0.220453468  -0.63601254   0.172145212
[5,] -0.05435431  -0.26235528   0.961759720   0.05127599  -0.024583093
```

So, $\boldsymbol{\Gamma} = [\boldsymbol{l}_1 \quad \cdots \quad \boldsymbol{l}_p]$ is the above matrix.

Sample principal components (Example: from Johnson and Wichern)

Transform the data: $\boldsymbol{W} = \boldsymbol{X\Gamma}$

```
> W = X %*% eigen(cov(track))$vectors
> W
           [,1]        [,2]         [,3]      [,4]        [,5]
 [1,] 10.543072 -10.527916  1.056659387 -5.700085  0.12072564
 [2,]  5.570539 -10.317952  0.579771227 -6.624057  0.08672935
 [3,]  6.705189  -9.590876 -0.128289543 -6.065719 -0.10773420
 [4,]  8.510143 -12.081228  0.662588769 -6.385425 -0.12840998
 [5,]  9.872154 -10.503207  0.358974840 -6.827804 -0.23907533
 [6,] 14.062902 -10.190444  0.823095907 -6.703009 -0.11981863
 [7,]  6.973711 -10.172400 -0.001861813 -6.238076 -0.07820397
 [8,] 12.096753 -12.332597 -0.381996516 -7.011537  0.04470289
 [9,] 12.277261  -7.378968  1.132888656 -6.780516  0.01945207
[10,]  7.684733  -8.635235  0.292771164 -6.062578 -0.06052581
[11,]  8.740515  -8.854721  0.296683219 -6.635868 -0.28168369
[12,]  5.806929 -11.211983  2.057016815 -6.965031 -0.11762634
[13,]  7.368046 -10.152644  0.754071151 -6.819313 -0.05356962
[14,] 11.263872 -11.344222  1.123849076 -5.461631 -0.21156989
```

Centred data

```
> MeanW <- matrix(rep(c(colMeans(W)),times=14),nrow=14,ncol=5,byrow=T)
> W - MeanW
           [,1]        [,2]        [,3]        [,4]         [,5]
 [1,]  1.4376565 -0.29260180  0.44050065  0.74853291  0.201197601
 [2,] -3.5348762 -0.08263869 -0.03638751 -0.17543886  0.167201310
 [3,] -2.4002270  0.64443800 -0.74444828  0.38289849 -0.027262233
 [4,] -0.5952725 -1.84591442  0.04643003  0.06319255 -0.047938015
 [5,]  0.7667385 -0.26789299 -0.25718390 -0.37918649 -0.158603360
 [6,]  4.9574860  0.04487018  0.20693717 -0.25439106 -0.039346660
 [7,] -2.1317042  0.06291364 -0.61802055  0.21054156  0.002267993
 [8,]  2.9913377 -2.09728323 -0.99815525 -0.56291943  0.125174851
 [9,]  3.1718449  2.85634591  0.51672992 -0.33189825  0.099924038
[10,] -1.4206830  1.60007856 -0.32338757  0.38604008  0.019946152
[11,] -0.3649005  1.38059323 -0.31947552 -0.18724994 -0.201211729
[12,] -3.2984865 -0.97666914  1.44085808 -0.51641343 -0.037154372
[13,] -1.7373697  0.08266930  0.13791241 -0.37069545  0.026902349
[14,]  2.1584560 -1.10890855  0.50769034  0.98698731 -0.131097925
```

Sample principal components (Example: from Johnson and Wichern)

Verification: the transformed data has sample mean $\bar{\boldsymbol{X}}\boldsymbol{\Gamma}$ and sample variance-covariance $\boldsymbol{\Gamma}'\boldsymbol{S}\boldsymbol{\Gamma}$

```
> colMeans(W)
[1]   9.10541567   -10.23531376   0.61615874   -6.44861798   -0.08047197
> cov(W)
             [,1]          [,2]          [,3]          [,4]          [,5]
[1,]  6.931074e+00  1.230186e-15 -3.067959e-16 -3.959973e-15 -1.553749e-15
[2,]  1.230186e-15  1.785144e+00 -2.053145e-16  5.961110e-16 -8.111907e-16
[3,] -3.067959e-16 -2.053145e-16  3.896499e-01  2.125370e-16  5.289134e-17
[4,] -3.959973e-15  5.961110e-16  2.125370e-16  2.295289e-01 -9.576007e-17
[5,] -1.553749e-15 -8.111907e-16  5.289134e-17 -9.576007e-17  1.415498e-02

> colMeans(X) %*% eigen(cov(track))$vectors
[1,] 9.105416     -10.23531     0.6161587   -6.448618   -0.08047197
> t(eigen(cov(track))$vectors) %*% cov(X) %*% eigen(cov(track))$vectors
             [,1]          [,2]          [,3]          [,4]          [,5]
[1,]  6.931074e+00 -4.579670e-16  0.000000e+00 -4.034967e-15 -1.172673e-15
[2,]  6.938894e-18  1.785144e+00  1.110223e-16  4.961309e-16 -8.448103e-16
[3,]  4.510281e-17  1.249001e-16  3.896499e-01  7.979728e-17  5.030698e-17
[4,] -3.945086e-15  6.331741e-16  8.153200e-17  2.295289e-01 -8.673617e-19
[5,] -1.648848e-15 -8.942906e-16  2.265983e-17 -2.928023e-17  1.415498e-02
```

Sample principal components (Example: from Johnson and Wichern)

What is the covariance between the principal components Y and X ?

```
> CovYX <-t(eigen(cov(track))$vectors) %*% cov(track)
> CovYX
           popul    education   employment       health    homevalue
[1,]  5.414610648  2.118472684  2.318086498  2.952692481 -0.3767337219
[2,] -0.126516450 -1.363623156  0.148002541  1.034416150 -0.4683420493
[3,]  0.001424797 -0.063052152  0.005782797  0.085899676  0.3747495991
[4,]  0.124338125 -0.125047208  0.011709729 -0.145983270  0.0117693216
[5,]  0.004275366  0.000131353 -0.013266834  0.002436713 -0.0003479733
```

Sample principal components (Example: from Johnson and Wichern)

What is the correlation between the principal components *Y* and *X* ?

```
> DiagvarY <- diag(1/sqrt(eigen(cov(track))$values))
> DiagvarY
          [,1]      [,2]     [,3]     [,4]     [,5]
[1,] 0.3798392 0.0000000 0.000000 0.000000 0.000000
[2,] 0.0000000 0.7484509 0.000000 0.000000 0.000000
[3,] 0.0000000 0.0000000 1.602001 0.000000 0.000000
[4,] 0.0000000 0.0000000 0.000000 2.087283 0.000000
[5,] 0.0000000 0.0000000 0.000000 0.000000 8.405147

> DiagSigma <- diag(1/sqrt(diag(cov(track))))
> DiagSigma
          [,1]      [,2]     [,3]      [,4]     [,5]
[1,] 0.4818197 0.0000000 0.000000 0.0000000 0.000000
[2,] 0.0000000 0.7521833 0.000000 0.0000000 0.000000
[3,] 0.0000000 0.0000000 1.117567 0.0000000 0.000000
[4,] 0.0000000 0.0000000 0.000000 0.7125655 0.000000
[5,] 0.0000000 0.0000000 0.000000 0.0000000 1.408059

> DiagvarY %*% CovYX %*% DiagSigma
            popul     education   employment     health    homevalue
[1,]  0.990949503  0.6052659980  0.98401791  0.7991766 -0.201490798
[2,] -0.045624159 -0.7676820085  0.12379586  0.5516751 -0.493568528
[3,]  0.001099766 -0.0759777242  0.01035319  0.0980571  0.845327227
[4,]  0.125046101 -0.1963265169  0.02731504 -0.2171247  0.034590250
[5,]  0.017314231  0.0008304416 -0.12461957  0.0145940 -0.004118245
```

Sample principal components (Example: from Johnson and Wichern)

Using correlation (Note the differences between using correlation and covariance)

```
> eigen(cor(track))
$values
[1] 3.02889606 1.29113796 0.57245566 0.09539848 0.01211184

$vectors
           [,1]         [,2]         [,3]        [,4]         [,5]
[1,] -0.5583589 -0.131392987  0.007945807 -0.55055321  0.606464575
[2,] -0.3132830 -0.628872546 -0.549030533  0.45265380 -0.006564747
[3,] -0.5682577 -0.004262264  0.117280380 -0.26811649 -0.769040874
[4,] -0.4866246  0.309560576  0.454923806  0.64798227  0.201325679
[5,]  0.1742664 -0.701005911  0.691224986 -0.01510711 -0.014203097
```

Sample principal components (Example: from Johnson and Wichern)

Using R function: princomp (with cov)

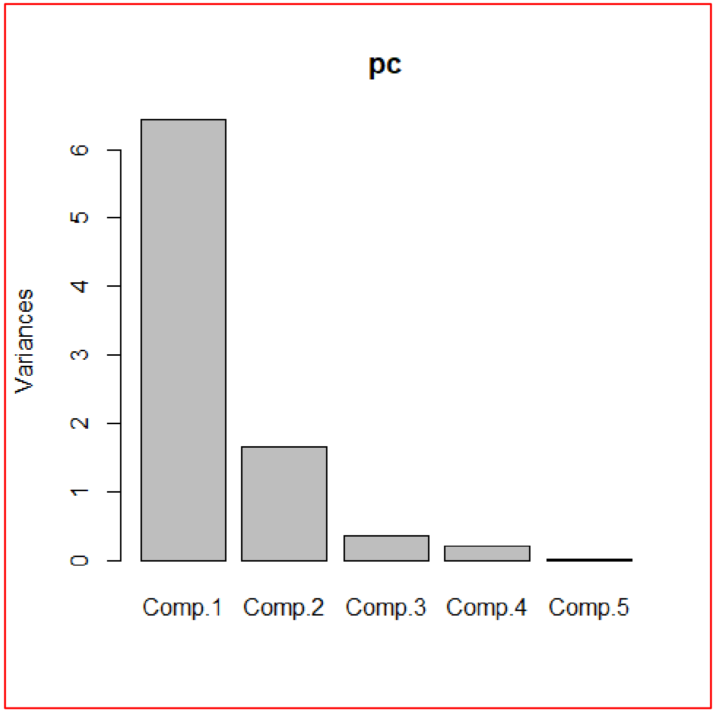
```
> pc <- princomp(track)
> pc
Call:
princomp(x = track)

Standard deviations:
   Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
2.5369267 1.2874914 0.6015129 0.4616644 0.1146469

 5  variables and  14 observations.
> summary(pc)
Importance of components:
```

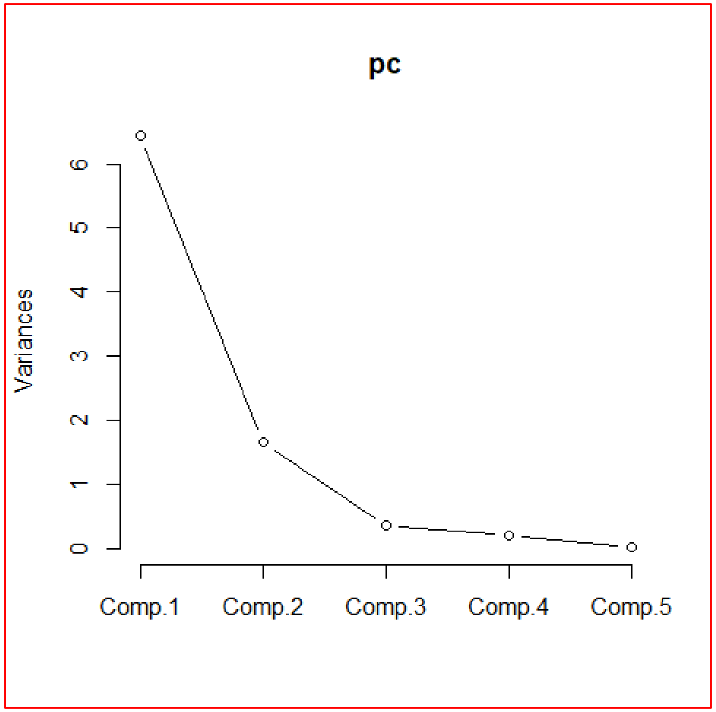
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Standard deviation	2.5369267	1.2874914	0.60151291	0.46166437	0.114646905
Proportion of Variance	0.7413268	0.1909337	0.04167579	0.02454972	0.001513975
Cumulative Proportion	0.7413268	0.9322605	0.97393630	0.99848603	1.000000000

Sample principal components (Example: from Johnson and Wichern)

Plots

```
> plot(pc)
> plot(pc, type="l")
```

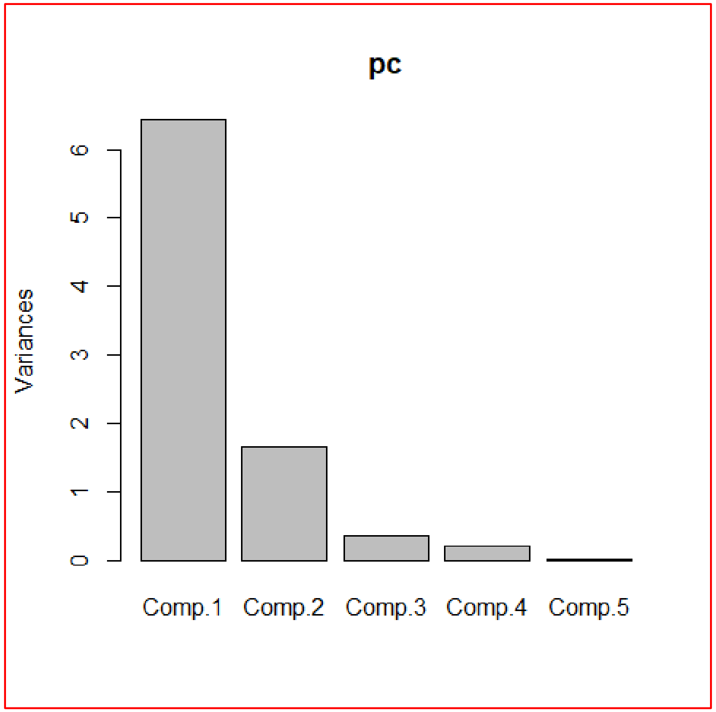


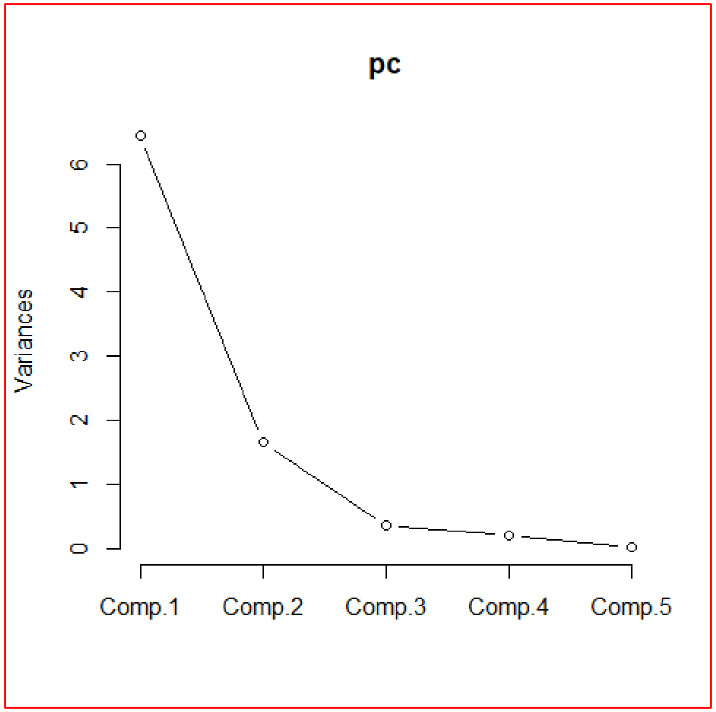


Sample principal components (Example: from Johnson and Wichern)

Plots

```
> plot(pc)
> plot(pc, type="l")
```





Sample principal components (Example: from Johnson and Wichern)

Loadings

```
> pc$loadings
```

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
popul	0.781			-0.542	0.302
education	0.306	-0.764	-0.162	0.545	
employment	0.334				-0.937
health	0.426	0.579	0.220	0.636	0.172
homevalue		-0.262	0.962		

```
> print(pc$loadings,digits=4,cutoff=0.001)
```

Loadings:

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
popul	0.7812	-0.0709	0.0037	-0.5417	0.3020
education	0.3056	-0.7639	-0.1618	0.5448	0.0093
employment	0.3344	0.0829	0.0148	-0.0510	-0.9373
health	0.4260	0.5795	0.2205	0.6360	0.1721
homevalue	-0.0544	-0.2624	0.9618	-0.0513	-0.0246

Sample principal components (Example: from Johnson and Wichern)

Scores

```
> pc$scores
        Comp.1      Comp.2      Comp.3      Comp.4       Comp.5
 [1,]  1.4376565 -0.29260180  0.44050065 -0.74853291  0.201197601
 [2,] -3.5348762 -0.08263869 -0.03638751  0.17543886  0.167201310
 [3,] -2.4002270  0.64443800 -0.74444828 -0.38289849 -0.027262233
 [4,] -0.5952725 -1.84591442  0.04643003 -0.06319255 -0.047938015
 [5,]  0.7667385 -0.26789299 -0.25718390  0.37918649 -0.158603360
 [6,]  4.9574860  0.04487018  0.20693717  0.25439106 -0.039346660
 [7,] -2.1317042  0.06291364 -0.61802055 -0.21054156  0.002267993
 [8,]  2.9913377 -2.09728323 -0.99815525  0.56291943  0.125174851
 [9,]  3.1718449  2.85634591  0.51672992  0.33189825  0.099924038
[10,] -1.4206830  1.60007856 -0.32338757 -0.38604008  0.019946152
[11,] -0.3649005  1.38059323 -0.31947552  0.18724994 -0.201211729
[12,] -3.2984865 -0.97666914  1.44085808  0.51641343 -0.037154372
[13,] -1.7373697  0.08266930  0.13791241  0.37069545  0.026902349
[14,]  2.1584560 -1.10890855  0.50769034 -0.98698731 -0.131097925
```

Sample principal components (Example: from Johnson and Wichern)

Using R function: princomp (with corr)

```
> pccorr <- princomp(track,cor=T)
> pccorr
Call:
princomp(x = track, cor = T)

Standard deviations:
   Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
1.7403724 1.1362825 0.7566080 0.3088664 0.1100538

 5  variables and  14 observations.
> summary(pccorr)
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4      Comp.5
Standard deviation     1.7403724 1.1362825 0.7566080 0.3088664 0.110053797
Proportion of Variance 0.6057792 0.2582276 0.1144911 0.0190797 0.002422368
Cumulative Proportion  0.6057792 0.8640068 0.9784979 0.9975776 1.000000000
```