L07 Observed principal components and their applications

1. Parameters related to principal components

   (1) For random vectors
   $X \in R^p$; standardized X, $Z \in R^p$; principal component vector of X, Y_x; principal component vector of Z, Y_z. has PC vector Y_x and standardized vector Z. Z has PC vector Y_z.
   With variance-covariance matrix $\Sigma = \mathrm{Cov}(X)$, variance matrix $V = \mathrm{diag}(\Sigma)$, and correlation matrix $\rho = V^{-1/2}\Sigma V^{-1/2}$, let $\Sigma = P_x\Lambda_x P_x'$ and $\rho = P_z\Lambda_z P_z'$ be the EVDs. Then

$$\begin{aligned}
&X \sim (\mu,\, \Sigma)\\
&Y_x = P_x'X \sim (P_x'\mu,\, P_x'\Sigma P_x) = (P_x'\mu,\, \Lambda_x)\\
&Z = V^{-1/2}(X-\mu) \sim (0,\, V^{-1/2}\Sigma V^{-1/2}) = (0,\, \rho)\\
&Y_z = P_z'Z \sim (0,\, P_z'\rho P_z) = (0,\, \Lambda_z)
\end{aligned}$$

   (2) Covariance matrices
   With four vectors X, Z, Y_x and Y_z, there are covariance matrices for 6 pairs. Each matrix may have multiple but equivalent expressions. For example

$$\mathrm{Cov}(Y_x,\, Z) = \mathrm{Cov}(P_x'X,\, V^{-1/2}(X-\mu)) = P_x'\Sigma V^{-1/2} = \Lambda_x P_x'V^{-1/2} = P_x'V^{1/2}\rho.$$

   (3) Correlation matrices
   There are also correlation matrices for 6 pairs of vectors. For example

$$\rho(Y_x,\, Z) = \Lambda_x^{-1/2}\mathrm{Cov}(Y_x,\, Z)\, I = \Lambda_x^{-1/2}\Lambda_x P_x'V^{-1/2} = \Lambda_x^{1/2}P_x'V^{-1/2}.$$

Ex1: In 8.3 p471 $X \sim (\mu,\, \Sigma)$ with $\Sigma = \begin{pmatrix} 2 & 0 & 0\\ 0 & 4 & 0\\ 0 & 0 & 4\end{pmatrix}$. Find $\rho(Y_x, Z)$.

$$\Sigma = \begin{pmatrix} 2 & 0 & 0\\ 0 & 4 & 0\\ 0 & 0 & 4\end{pmatrix} = V = P_x\Lambda_x P_x' = \begin{pmatrix} 0 & 0 & 1\\ 0 & 1 & 0\\ 1 & 0 & 0\end{pmatrix}\begin{pmatrix} 4 & 0 & 0\\ 0 & 4 & 0\\ 0 & 0 & 2\end{pmatrix}\begin{pmatrix} 0 & 0 & 1\\ 0 & 1 & 0\\ 1 & 0 & 0\end{pmatrix}.$$

$$\rho(Y_x, Z) = \Lambda_x^{1/2}P_x'V^{-1/2} = \begin{pmatrix} 2 & 0 & 0\\ 0 & 2 & 0\\ 0 & 0 & \sqrt{2}\end{pmatrix}\begin{pmatrix} 0 & 0 & 1\\ 0 & 1 & 0\\ 1 & 0 & 0\end{pmatrix}\begin{pmatrix} \frac{1}{\sqrt{2}} & 0 & 0\\ 0 & \frac{1}{2} & 0\\ 0 & 0 & \frac{1}{2}\end{pmatrix} = \begin{pmatrix} 0 & 0 & 1\\ 0 & 1 & 0\\ 1 & 0 & 0\end{pmatrix}$$

2. Observed principal components

   (1) Population principal components
   By entering population covariance matrix Σ into SAS and do the procedures

`proc princomp cov;`
 `var x1 x2 x3;`
 `run;`
	`proc princomp;`
	 `var x1 x2 x3;`
	 `run;`

   one can get the EVDs for Σ and ρ from which the inference on the PCs of population and standardized population can be made.

   (2) Entering data from file into SAS
   But in practice Σ is most likely unavailable. In stead, we have sample from the population. For example file sample.txt

`1 2 3 4 5`
 `6 7 8 9 10`
 `.......`
 `8 4 3 1`
	`data a;`
	 `infile "C:\sample.txt";`
	 `input y x1 x2 x3 x4;`

(3) Requesting EVDs of S and R
SAS can calculate S and R from sample and do EVDs of S and R to get the estimated PCs for population and standardized population

`proc princomp cov;`
`var x1 x2 x3 x4;`
`run;`
	`proc princomp;`
	`var x1 x2 x3 x4;`
	`run;`

(4) Observed principal components
The PCs, for example $Y_1 = \frac{1}{\sqrt{2}}X_1 - \frac{1}{\sqrt{2}}X_2$, is based on the EVD of S, or simply based on n observations from population. With n observations on X_1 and X_2, there are n observations on Y_1. Generally, there are n observations on the principal component vector.

`proc princomp cov out=b;`
`var x1 x2 x3 x4;`
`run;`
	`proc print;`
	`var prin1 prin2 prin3 pin4;`
	`run;`

3. An example of the usage of principal components

(1) Regression
Regression model $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ relates y to two predictors x_1 and x_2. Suppose the observations on y and six candidate variables are stored in file Example.txt with order y, $z1$, $z2$, $z3$, $z4$, $z5$, $z6$. We may select first two predictors $z1$ and $z2$ as x_1 and x_2. We may also find observations on principal components and use the first two as x_1 and x_2.

(2) SAS
To implement our plan we execute SAS below.

```
data a;
   infile "D:\Example.txt";
   input y z1 z2 z3 z4 z5;
proc princomp cov out=b;
   var z1 z2 z3 z4 z5;
   run;
proc reg;
   model y=z1 z2;
   run;
proc reg;
   model y=prin1 prin2;
   run;
```

(3) Results
The first model produced the coefficient of determination $R^2 = 0.0993$ which means that about 10% of variation in y are explained by two predictor variable $z1$ and $z2$.
The second model produced $R^2 = 0.1424$ which means that about 14.24% of variation in y are explained by two principal components.