Principal Component Analysis (I)

- Objectives
- Population principal components

(Materials from Johnson and Wichern, Applied Multivariate Statistical Analysis)

Objectives

- Dimension reduction through linear combinations
- An intermediate step in a more complex data analysis (eg. Cluster analysis)
- Easier to interpret?

Population principal components

Let $\boldsymbol{X} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{X} = (X_1, X_2, \ldots, Xp)'$ and let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ be the eigenvalues of $\boldsymbol{\Sigma}$.

Assume that $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Yp)'$ where Y_i is a linear combination of $\boldsymbol{X}$. That is

$$Y_i = \boldsymbol{l}_i' \boldsymbol{X}$$

Then, Var (Y_i) = $\boldsymbol{l}_i' \boldsymbol{\Sigma}\, \boldsymbol{l_i}$ and Cov (Y_i , Y_j) = $\boldsymbol{l}_i' \boldsymbol{\Sigma}\, \boldsymbol{l_j}$ $\qquad \boldsymbol{i \neq j}$

Population principal components

1^{st} $principal\ component$: $Y_1 = \boldsymbol{l}_1'\boldsymbol{X}$ that maximizes Var $(\boldsymbol{l}_1'\boldsymbol{X})$ subject to $\boldsymbol{l}_1'\ \boldsymbol{l}_1 = 1$

2^{nd} $principal\ component$: $Y_2 = \boldsymbol{l}_2'\boldsymbol{X}$ that maximizes Var $(\boldsymbol{l}_2'\boldsymbol{X})$ subject to $\boldsymbol{l}_2'\ \boldsymbol{l}_2 = 1$ and Cov $(\boldsymbol{l}_1'\boldsymbol{X}, \boldsymbol{l}_2'\boldsymbol{X})$ = 0

$$\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots$$

i^{th} $principal\ component$: $Y_i = \boldsymbol{l}_i'\boldsymbol{X}$ that maximizes Var $(\boldsymbol{l}_i'\boldsymbol{X})$ subject to $\boldsymbol{l}_i'\ \boldsymbol{l}_i = 1$ and Cov $(\boldsymbol{l}_i'\boldsymbol{X}, \boldsymbol{l}_k'\boldsymbol{X})$ = 0

for $k < i$

Population principal components

Main Result

Let $\boldsymbol{\Sigma}$ be the covariance matrix of $\boldsymbol{X} = (X_1, X_2, \dots, Xp)'$. Further, let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ be the eigenvalues of $\boldsymbol{\Sigma}$, with associated eigenvectors $\boldsymbol{e}_1$, $\boldsymbol{e}_2$,…, $\boldsymbol{e}_p$

1^{st} *principal component*: $Y_1 = \boldsymbol{e}_1'\boldsymbol{X}$

2^{nd} *principal component*: $Y_2 = \boldsymbol{e}_2'\boldsymbol{X}$

$$\begin{matrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{matrix}$$

i^{th} *principal component*: $Y_i = \boldsymbol{e}_i'\boldsymbol{X}$

Population principal components

<u>Proportion of total variance due to the i^{th} principal component</u>

Let $\boldsymbol{\Sigma}$ be the covariance matrix of $\boldsymbol{X} = (X_1, X_2, \dots, Xp)'$ with diagonal elements σ_{11}, σ_{22},…, σ_{pp},

Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ be the eigenvalues of $\boldsymbol{\Sigma}$, with associated eigenvectors $\boldsymbol{e}_1$, $\boldsymbol{e}_2$,…, $\boldsymbol{e}_p$

Define total variance $=\sum_{i=1}^{p} Var(X_i) = \sum_{i=1}^{p} \sigma_{ii}$

Can be shown that $\sum_{i=1}^{p} \sigma_{ii} = \sum_{i=1}^{p} \lambda_i$

Hence, proportion of total population variance due to the k^{th} principal component is

$$\boldsymbol{\frac{\lambda_k}{\sum_{i=1}^{p} \lambda_i}}$$

Population principal components

Let $\boldsymbol{X} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{X} = (X_1, X_2, \dots, Xp)'$. $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ are the eigenvalues of $\boldsymbol{\Sigma}$ with corresponding normalized eigenvectors be $\boldsymbol{e}_1$, $\boldsymbol{e}_2$,…, $\boldsymbol{e}_p$. Let $\boldsymbol{\Gamma} = (\boldsymbol{e}_1, \boldsymbol{e}_2, \dots, \boldsymbol{e}_p)$, then $\boldsymbol{Y} = \boldsymbol{\Gamma}'(\boldsymbol{X} - \boldsymbol{\mu})$ is the **principal component transformation** of $\boldsymbol{X}$.

Properties

1. $E(Y_i) = 0, \quad i = 1, \dots, p$
2. $Var(Y_i) = \lambda_i, \quad i = 1, \dots, p$
3. $Cov(Y_i, Y_j) = 0, \quad i \neq j$
4. $Var(Y_1) \geq Var(Y_2) \geq \cdots \geq Var(Y_p) \geq 0$
5. $\sum_{i=1}^{p} Var(Y_i) = trace(\boldsymbol{\Sigma})$
6. $\prod_{i=1}^{p} Var(Y_i) = |\boldsymbol{\Sigma}|$

Population principal components

<u>Correlation coefficients between a principal component and the original variables</u>

Let $\boldsymbol{X} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{X} = (X_1, X_2, \dots, Xp)'$ and $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ be the eigenvalues of $\boldsymbol{\Sigma}$, with associated eigenvectors $\boldsymbol{e}_1$, $\boldsymbol{e}_2$,…, $\boldsymbol{e}_p$. If $Y_1 = \boldsymbol{e}_1'\boldsymbol{X}, Y_2 = \boldsymbol{e}_2'\boldsymbol{X}, \dots, Y_p = \boldsymbol{e}_p'\boldsymbol{X}$ are the principal components, then

$$\rho_{Y_i, X_k} = \frac{e_{ki}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}$$

where $i, k = 1,2, \dots, p$ are the correlation coefficients between the components Y_i and the variables X_k.

Population principal components

Example

Let $\boldsymbol{X} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{X} = (X_1, X_2, X_3)'$ and $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix}$

The principal components are

$$Y_1 = \boldsymbol{e}_1'\boldsymbol{X} = -.383X_1 + .924X_2$$
$$Y_2 = \boldsymbol{e}_2'\boldsymbol{X} = X_3$$
$$Y_3 = \boldsymbol{e}_3'\boldsymbol{X} = .924X_1 + .383X_2$$

$$\begin{aligned} Var(Y_1) &= Var(-.383X_1 + .924X_2) \\ &= (-.383)^2 Var(X_1) + (.924)^2 Var(X_2) + 2(-.383)(.924)Cov(X_1, X_2) \\ &= 5.83 = \lambda_1 \end{aligned}$$

and so on…

```
> Sigma <- matrix(c(1,-2,0,-2,5,0,0,0,2),nrow=3)
> Sigma
     [,1] [,2] [,3]
[1,]    1   -2    0
[2,]   -2    5    0
[3,]    0    0    2
> eigen(Sigma)
$values
[1] 5.8284271 2.0000000 0.1715729

$vectors
           [,1] [,2]      [,3]
[1,] -0.3826834    0 0.9238795
[2,]  0.9238795    0 0.3826834
[3,]  0.0000000    1 0.0000000
```

```
> VarY <- t(eigen(Sigma)$vectors) %*% Sigma %*% eigen(Sigma)$vectors
> VarY
              [,1] [,2]          [,3]
[1,]  5.828427e+00    0 -8.881784e-16
[2,]  0.000000e+00    2  0.000000e+00
[3,] -7.979728e-16    0  1.715729e-01
>
```

Population principal components

<u>Example (Cont'd)</u> **What is the covariance between the principal components Y and X ?**

```
> CovYX <-t(eigen(Sigma)$vectors) %*% Sigma
> CovYX
          [,1]      [,2] [,3]
[1,] -2.2304425 5.3847645    0
[2,]  0.0000000 0.0000000    2
[3,]  0.1585127 0.0656581    0
```

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

```
> DiagvarY <- diag(1/sqrt(diag(VarY)))
> DiagvarY
          [,1]      [,2]     [,3]
[1,] 0.4142136 0.0000000 0.000000
[2,] 0.0000000 0.7071068 0.000000
[3,] 0.0000000 0.0000000 2.414214
> DiagSigma <- diag(1/sqrt(diag(Sigma)))
> DiagSigma
     [,1]      [,2]      [,3]
[1,]    1 0.0000000 0.0000000
[2,]    0 0.4472136 0.0000000
[3,]    0 0.0000000 0.7071068
> DiagvarY %*% CovYX %*% DiagSigma
           [,1]       [,2] [,3]
[1,] -0.9238795 0.99748421    0
[2,]  0.0000000 0.00000000    1
[3,]  0.3826834 0.07088902    0
```

[Correlations between the first PC and **X**]
[Correlations between the second PC and **X**]
[Correlations between the third PC and **X**]

Population principal components (**using the correlation matrix**)

Let $\boldsymbol{X} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{X} = (X_1, X_2, \ldots, Xp)'$ and $\boldsymbol{\Sigma} = \{\sigma_{ij}\}$.

Let $\boldsymbol{D} = \begin{bmatrix} \sigma_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sigma_{pp} \end{bmatrix}$ be a diagonal matrix with diagonal elements σ_{ii}, $i = 1, \ldots, p$.

Let $\boldsymbol{\rho}$ be the correlation matrix of $\boldsymbol{X}$. Then, $\boldsymbol{\rho} = (\boldsymbol{D}^{\mathbf{1/2}})^{-\mathbf{1}} \boldsymbol{\Sigma} (\boldsymbol{D}^{\mathbf{1/2}})^{-\mathbf{1}}$

Standardized variables: $Z_i = (X_i - \mu_i)/\sqrt{\sigma_{ii}}$

Consider $\boldsymbol{Z} = (\boldsymbol{D}^{\frac{1}{2}})^{-\mathbf{1}}(\boldsymbol{X} - \boldsymbol{\mu})$ where $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Zp)'$

Then, $\boldsymbol{Z} \sim (\mathbf{0}, \boldsymbol{\rho})$

Population principal components (**using the correlation matrix)**

Now, the standardized vector $\boldsymbol{Z} \sim (\mathbf{0}, \boldsymbol{\rho})$ and if $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ are the eigenvalues of $\boldsymbol{\rho}$, with associated eigenvectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_p$

1^{st} *principal component*: $Y_1 = \boldsymbol{e}_1'\boldsymbol{Z}$
2^{nd} *principal component*: $Y_2 = \boldsymbol{e}_2'\boldsymbol{Z}$

. . .

. . .

. . .

p^{th} *principal component*: $Y_p = \boldsymbol{e}_p'\boldsymbol{Z}$

total variance $= \sum_{i=1}^{p} Var(Y_i) = \sum_{i=1}^{p} Var(Z_i) = p$

Hence, proportion of total population variance due to the k^{th} principal component is $\boldsymbol{\frac{\lambda_k}{p}}$

Correlation between principal component Y_i and Z_k is $\rho_{Y_i, X_k = e_{ki\sqrt{\lambda_i}}}$

Population principal components (**using the correlation matrix)**

Example

Let $\boldsymbol{X} \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ where $\boldsymbol{X} = (X_1, X_2, X_3)'$ and $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & -2 & 0 \\ -2 & 5 & 0 \\ 0 & 0 & 2 \end{pmatrix}$

The covariance matrix is $\boldsymbol{\rho}$ given in the right-hand box (refer to R-output).

The principal components using the covariance matrix are

$$Y_1 = \boldsymbol{e}_1'\boldsymbol{Z} = -0.7071068Z_1 + .7071068Z_2$$
$$Y_2 = \boldsymbol{e}_2'\boldsymbol{Z} = Z_3$$
$$Y_3 = \boldsymbol{e}_3'\boldsymbol{Z} = .7071068Z_1 + .7071068Z_2$$

Variance-covariance matrix of **Y**

```
> VarY <- t(eigen(rho)$vectors) %*% rho %*% eigen(rho)$vectors
> VarY
          [,1]                [,2]           [,3]
[1,]  1.894427e+00            0              -4.440892e-16
[2,]  0.000000e+00            1              0.000000e+00
[3,] -4.579670e-16            0              1.055728e-01
```

```
> D <- diag(diag(Sigma))
> D
     [,1]  [,2]  [,3]
[1,]   1    0    0
[2,]   0    5    0
[3,]   0    0    2
> Droot <- diag(sqrt(diag(Sigma)))
> Droot
              [,1]         [,2]         [,3]
[1,]          1            0.000000     0.000000
[2,]          0            2.236068     0.000000
[3,]          0            0.000000     1.414214
> rho <- solve(Droot) %*% Sigma %*%  solve(Droot)
> rho
              [,1]         [,2]         [,3]
[1,]          1.0000000    -0.8944272   0
[2,]          -0.8944272   1.0000000    0
[3,]          0.0000000    0.0000000    1
> eigen(rho)
$values
[1] 1.8944272   1.0000000   0.1055728

$vectors
              [,1]         [,2]         [,3]
[1,]          -0.7071068   0            0.7071068
[2,]          0.7071068    0            0.7071068
[3,]          0.0000000    1            0.0000000
```

Population principal components (<u>**using the correlation matrix)**</u>

<u>**Example (Cont'd)**</u> **What is the covariance between the principal components Y and Z ?**

```
> CovYZ <-t(eigen(rho)$vectors) %*% rho
> CovYZ
           [,1]        [,2]  [,3]
[1,] -1.33956231  1.33956231  0
[2,]  0.00000000  0.00000000  1
[3,]  0.07465125  0.07465125  0
```

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

```
> DiagvarY <- diag(1/sqrt(diag(VarY)))
> DiagvarY
          [,1]       [,2]      [,3]
[1,] 0.7265425  0.0000000  0.000000
[2,] 0.0000000  1          0.000000
[3,] 0.0000000  0.0000000  3.077684
> Diagrho <- diag(1/sqrt(diag(rho)))
> Diagrho
     [,1]  [,2]  [,3]
[1,]  1     0     0
[2,]  0     1     0
[3,]  0     0     1
> DiagvarY %*% CovYZ %*% Diagrho
           [,1]       [,2]  [,3]
[1,] -0.9732490  0.9732490  0     [Correlations between the first PC and Z]
[2,]  0.0000000  0.0000000  1     [Correlations between the second PC and Z]
[3,]  0.2297529  0.2297529  0     [Correlations between the third PC and Z]
```