# Tut 8: PCA Dimensional Reduction

Jan 2023

When variances $\mathrm{Var}(x_{\cdot j})$ for features/columns $x_{\cdot j}$ differ a lot, we need to perform scaling:

`pca$scale`: $\sqrt{\frac{\sum_i (x_{ij}-\overline{x}_{\cdot j})^2}{n-1}}$

However, you do not need to scale the data unless it is stated in the question.

---

Original data: $X$; Data shifted to centre: $\widetilde{X}$

`pca$center`: $\overline{x}_{\cdot j}$

`pca$sdev`: $\sqrt{\lambda_i}$

`pca$rotation`: $[\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots]$

`pca$x`: $[\widetilde{X}\boldsymbol{e}_1, \widetilde{X}\boldsymbol{e}_2, \cdots]$

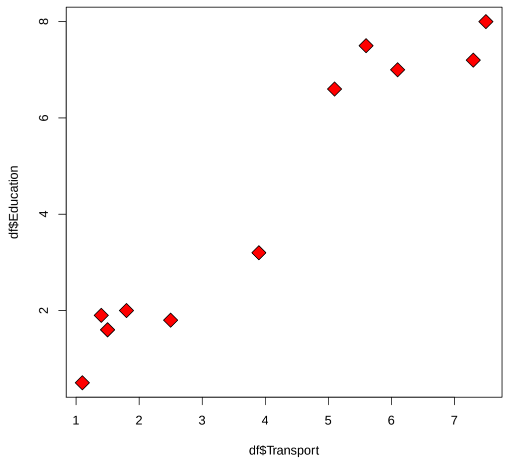
1. You are given 12 communities that were rated according to transportation and education — the higher the score the better. For example, a better transportation system will score higher. Higher education facilities will score higher as well. The table below shows the score for 12 communities in the two criteria:

| Obs | Transportation | Education |
|---|---|---|
| 1 | 1.1 | 0.5 |
| 2 | 3.9 | 3.2 |
| 3 | 1.5 | 1.6 |
| 4 | 5.6 | 7.5 |
| 5 | 2.5 | 1.8 |
| 6 | 7.3 | 7.2 |
| 7 | 1.4 | 1.9 |
| 8 | 6.1 | 7.0 |
| 9 | 1.5 | 1.6 |
| 10 | 5.1 | 6.6 |
| 11 | 1.8 | 2.0 |
| 12 | 7.5 | 8.0 |

(a) Plot a scatterplot to visualize your data.

*Solution.* A simple R script:

```
d.f = data.frame(
        Transport = c(1.1,3.9,1.5,5.6,2.5,7.3,1.4,6.1,1.5,5.1,1.8,7.5),
        Education = c(0.5,3.2,1.6,7.5,1.8,7.2,1.9,7.0,1.6,6.6,2.0,8.0)
)
plot(d.f$Transport,d.f$Education,type='p',pch=23,bg="red",cex=2)
```




□

(b) Generate two principal components for the data.

*Solution.* Calculating using R script:

```
Transport = c(1.1,3.9,1.5,5.6,2.5,7.3,1.4,6.1,1.5,5.1,1.8,7.5)
Education = c(0.5,3.2,1.6,7.5,1.8,7.2,1.9,7.0,1.6,6.6,2.0,8.0)
X = data.frame(Transport, Education)
PC = prcomp(X)
print(PC)
```

```
Standard deviations:
[1] 3.7504618 0.4861164

Rotation:
                PC1        PC2
Transport 0.6429319 -0.7659234
Education 0.7659234  0.6429319
```

Manual calculation:

i. Shift $\boldsymbol{X}$ to centre, i.e. find $\mu_1 = 3.775$, $\mu_2 = 4.075$ and generate table $\boldsymbol{X}^*$ below.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| $x_1$ | -2.675 | 0.125 | -2.275 | 1.825 | -1.275 | 3.525 | -2.375 |
| | | | 2.325 | -2.275 | 1.325 | -1.975 | 3.725 |
| $x_2$ | -3.575 | -0.875 | -2.475 | 3.425 | -2.275 | 3.125 | -2.175 |
| | | | 2.925 | -2.475 | 2.525 | -2.075 | 3.925 |

ii. Calculate the covariance matrix for $\boldsymbol{X}^*$, i.e.

$$C = \frac{1}{12-1}(\boldsymbol{X}^*)^T\boldsymbol{X}^* = \begin{bmatrix} 5.952955 & 6.810227 \\ 6.810227 & 8.349318 \end{bmatrix}$$

iii. Find the eigenvalues and eigenvectors of $C$ which characterises the "variance" of the data $\boldsymbol{X}^*$, i.e.

$$|C - \lambda I| = (5.952955 - \lambda)(8.349318 - \lambda) - 6.810227^2 = \lambda^2 - 14.302273\lambda + 3.323923 = 0$$

Using calculator, we obtain

$$\lambda = 0.236310,\ 14.065963 =: \lambda_1, \lambda_2$$

iv. We then find the eigenvalues for $\lambda_1$ and $\lambda_2$:

$$\boldsymbol{e}_1 = \frac{1}{\sqrt{6.810227^2 + (-5.716645)^2}} \begin{bmatrix} 6.810227 \\ -(5.952955 - 0.236310) \end{bmatrix} = \begin{bmatrix} 0.765923 \\ -0.6429319 \end{bmatrix}$$

$$\boldsymbol{e}_2 = \frac{1}{\sqrt{6.810227^2 + (8.113008)^2}} \begin{bmatrix} 6.810227 \\ -(5.952955 - 14.065963) \end{bmatrix} = \begin{bmatrix} 0.6429319 \\ 0.765923 \end{bmatrix}$$

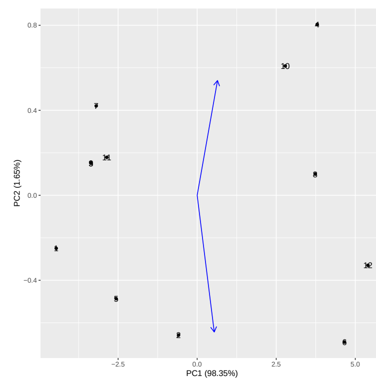
Observe that when $\boldsymbol{e}_1 = [a, b]$ and $\boldsymbol{e}_1 \cdot \boldsymbol{e}_2 = 0$, $\boldsymbol{e}_2 = [b, -a]$ is an answer.

v. Calculate the "principal components":

$PC_1 = \sum_{i=1}^{2} e_{i1}(X_i - \mathbb{E}(X_i)) = 0.6429319x_1^* + 0.7659234x_2^*$

$PC_2 = \sum_{i=1}^{2} e_{i2}(X_i - \mathbb{E}(X_i)) = 0.7659234x_1^* - 0.6429319x_2^*$

| $PC_1$ | $PC_2$ |
|---|---|
| -4.4580188 | -0.24963649 |
| -0.5898165 | -0.65830582 |
| -3.3583304 | 0.15121924 |
| 3.7966382 | 0.80423156 |
| -2.5622138 | -0.48611775 |
| 4.6598454 | -0.69071771 |
| -3.1928465 | 0.42069114 |
| 3.7351425 | 0.09980394 |
| -3.3583304 | 0.15121924 |
| 2.7858412 | 0.60855455 |
| -2.8590814 | 0.17861498 |
| 5.4011705 | -0.32955688 |




By rotating the "principal components" and shift it to the centre $(\mu_1, \mu_2)$, we can "recover" the original data. □

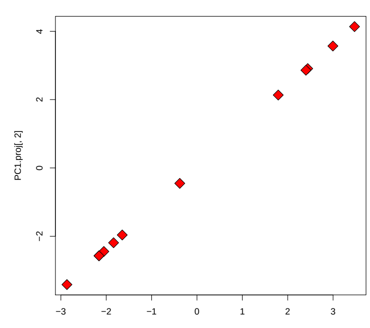
(c) Choose one suitable principal component to represent the data.

*Solution.* It must be the first principal component, i.e. $PC_1$. □

(d) Plot your data with the principal component you chose in (c).

*Solution.* Projecting the centred data $\boldsymbol{X}^*$ to the space span by PC1:

| $x_1^\#$ | $x_2^\#$ |
|---|---|
| -2.8662 | -3.4145 |
| -0.3792 | -0.4518 |
| -2.1592 | -2.5722 |
| 2.4410 | 2.9079 |
| -1.6473 | -1.9625 |
| 2.9960 | 3.5691 |
| -2.0528 | -2.4455 |
| 2.4014 | 2.8608 |
| -2.1592 | -2.5722 |
| 1.7911 | 2.1337 |
| -1.8382 | -2.1898 |
| 3.4726 | 4.1369 |




□

(e) With the eigenvalues computed in (b), calculate the proportion of variance explained by each component and the cumulative proportion.

*Solution.* `print(summary(PC))`

```
                         PC1     PC2
Standard deviation     3.7505 0.48612
Proportion of Variance 0.9835 0.01652
Cumulative Proportion  0.9835 1.00000
```

Manual calculation:

| | Eigenvalue | PVE | Cumulative PVE |
|---|---|---|---|
| PC1 | 14.0660 | $\frac{14.0660}{14.3023} = 0.9835$ | 0.9835 |
| PC2 | 0.2363 | $\frac{0.2363x}{14.3023} = 0.0165$ | 1 |
| $\lambda_1 + \lambda_2$ | 14.3023 | | |

□

(f) With a targeted explained variation of 95%, how many principal components should be considered? State the total variation explained.

*Solution.* One principal component, PC1. Total variance explained is 98.35%. □

2. (May 2020 Final Q4(a)) Given the following data with 8 observations in Table 4.1:

Table 4.1: Data with 2 features.

| Obs | x | y |
|---|---|---|
| A | 5.51 | 5.35 |
| B | 20.82 | 24.03 |
| C | -0.77 | -0.57 |
| D | 19.30 | 19.39 |
| E | 14.24 | 12.77 |
| F | 9.74 | 9.68 |
| G | 11.59 | 12.06 |
| H | -6.08 | -5.22 |

Find the first principle component and project the data $(5.51, 5.35)$ to the space span by the first principal component. (4 marks)

*Solution.* First, we need to find the mean: $\overline{x} = 9.29375, \quad \overline{y} = 9.68625$ ............. [0.5 mark]

and shift the data to centre at the mean:

$X=$

| Obs | x | y |
|---|---|---|
| A | -3.78375 | -4.33625 |
| B | 11.52625 | 14.34375 |
| C | -10.06375 | -10.25625 |
| D | 10.00625 | 9.70375 |
| E | 4.94625 | 3.08375 |
| F | 0.44625 | -0.00625 |
| G | 2.29625 | 2.37375 |
| H | -15.37375 | -14.90625 |

.................................................................................. [0.5 mark]

Form the covariant matrix and

$$\frac{1}{8-1} X^T X = \begin{bmatrix} 614.8648 & 631.9173 \\ 631.9173 & 661.2402 \end{bmatrix} = \begin{bmatrix} 87.83783 & 90.27390 \\ 90.27390 & 94.46288 \end{bmatrix} \quad \text{[0.5 mark]}$$

By solving the eigenvalue problem

$$\begin{vmatrix} 87.83783 - \lambda & 90.27390 \\ 90.27390 & 94.46288 - \lambda \end{vmatrix} = \lambda^2 - 182.3007\lambda + 148.0374 = 0 \qquad [1 \text{ mark}]$$

leads to the eigenvalues 181.4850, 0.8157

The first principle component corresponds $\boldsymbol{v}$ to the linear algebra problem of the eigenvalue 181.4850

$$\begin{bmatrix} 87.83783 - 181.4850 & 90.27390 \\ 90.27390 & 94.46288 - 181.4850 \end{bmatrix} \boldsymbol{v} = \boldsymbol{0}$$

i.e.

$$\boldsymbol{v} = \frac{1}{\sqrt{90.27390^2 + 93.64717^2}} \begin{bmatrix} 90.27390 \\ 93.64717 \end{bmatrix} = \begin{bmatrix} 0.69402 \\ 0.71995 \end{bmatrix} \qquad [0.5 \text{ mark}]$$

The projection of (5.51, 5.35) to the first principle component space is

$$(-3.78375, -4.33625) \cdot (0.69402, 0.71995) \begin{bmatrix} 0.69402 \\ 0.71995 \end{bmatrix} + \begin{bmatrix} 9.29375 \\ 9.68625 \end{bmatrix} = \begin{bmatrix} 5.3046 \\ 5.5481 \end{bmatrix} \qquad [1 \text{ mark}]$$

□

3. (Jan 2021 Final Q3(a)) Given the following data with 11 observations in Table 3.1:

Table 3.1: Data with two features.

| Obs | x | y |
|---|---|---|
| 1 | -5.79 | 4.91 |
| 2 | -3.73 | 4.87 |
| 3 | -3.25 | 3.98 |
| 4 | -2.61 | 4.09 |
| 5 | -2.76 | 4.90 |
| 6 | 2.81 | -5.34 |
| 7 | 2.92 | -6.15 |
| 8 | 1.97 | -4.51 |
| 9 | 5.17 | -5.29 |
| 10 | 2.66 | -7.10 |
| 11 | 3.47 | -4.70 |

Find the proportions of variance and the principle components. (5 marks)

*Solution.* First, we need to find the mean: $\overline{x} = 0.07818182$, $\quad \overline{y} = -0.94$ . . . . . . . . . . . [0.5 mark]

and shift the data to centre at the mean:

$X=$

| Obs | x | y |
|---|---|---|
| 1 | -5.868182 | 5.85 |
| 2 | -3.808182 | 5.81 |
| 3 | -3.328182 | 4.92 |
| 4 | -2.688182 | 5.03 |
| 5 | -2.838182 | 5.84 |
| 6 | 2.731818 | -4.40 |
| 7 | 2.841818 | -5.21 |
| 8 | 1.891818 | -3.57 |
| 9 | 5.091818 | -4.35 |
| 10 | 2.581818 | -6.16 |
| 11 | 3.391818 | -3.76 |

. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . [0.5 mark]

Form the covariant matrix and

$$\frac{1}{11-1}X^T X = \begin{bmatrix} 138.5108 & -187.3119 \\ -187.3119 & 281.8462 \end{bmatrix} = \begin{bmatrix} 13.85108 & -18.73119 \\ -18.73119 & 28.18462 \end{bmatrix}.$$ [1 mark]

By solving the eigenvalue problem

$$\begin{vmatrix} 13.85108-\lambda & -18.73119 \\ -18.73119 & 28.18462-\lambda \end{vmatrix} = \lambda^2 - 42.0357\lambda + 39.52995 = 0$$

leads to the eigenvalues 0.962425, 41.073275 . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . [1 mark]

The proportions of variance are

$$\frac{41.073275}{41.073275+0.962425} = 0.977105, \quad \frac{0.962425}{41.073275+0.962425} = 0.022895$$ [0.5 mark]

The first principle component corresponds $\boldsymbol{v}$ to the linear algebra problem of the eigenvalue 41.073275

$$\begin{bmatrix} 13.85108 - 41.073275 & -18.73119 \\ -18.73119 & 28.18462 - 41.073275 \end{bmatrix} \boldsymbol{v} = \boldsymbol{0}$$

i.e.

$$\boldsymbol{v} = \frac{1}{\sqrt{(-18.73119)^2 + (27.222195)^2}} \begin{bmatrix} -18.73119 \\ 27.222195 \end{bmatrix} = \begin{bmatrix} -0.566856 \\ 0.823817 \end{bmatrix}$$ [1 mark]

The second principle component is orthogonal to the first principle component:

$$\begin{bmatrix} 0.823817 \\ 0.566856 \end{bmatrix}$$ [0.5 mark]

□