

Tut 8: PCA Dimensional Reduction

Jan 2022

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

$$\text{pca}\$scale: \sqrt{\frac{\sum_i (x_{ij} - \bar{x}_{.j})^2}{n-1}}$$

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

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

$\text{pca}\$center: \bar{x}_{.j}$

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

$\text{pca}\$rotation: [e_1, e_2, \dots]$

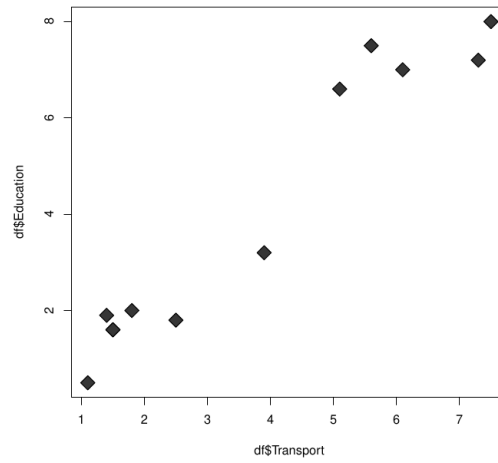
$\text{pca}\$x: [\tilde{X}e_1, \tilde{X}e_2, \dots]$

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:

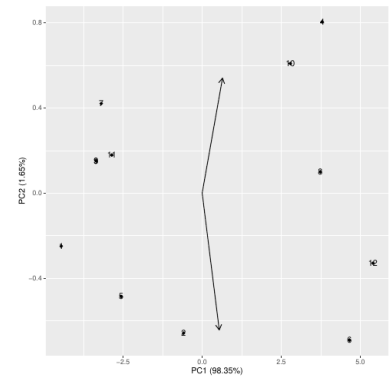
Obs	Transportation	Education
1	1.1	0.5
2	3.9	3.2
3	1.5	1.6
4	5.6	7.5
5	2.5	1.8
6	7.3	7.2
7	1.4	1.9
8	6.1	7.0
9	1.5	1.6
10	5.1	6.6
11	1.8	2.0
12	7.5	8.0

- (a) Plot a scatterplot to visualize your data.



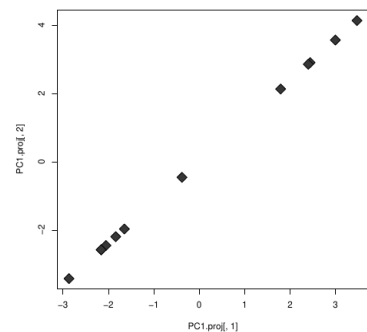


(b) Generate two principal components for the data.



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

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



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

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

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)

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)

