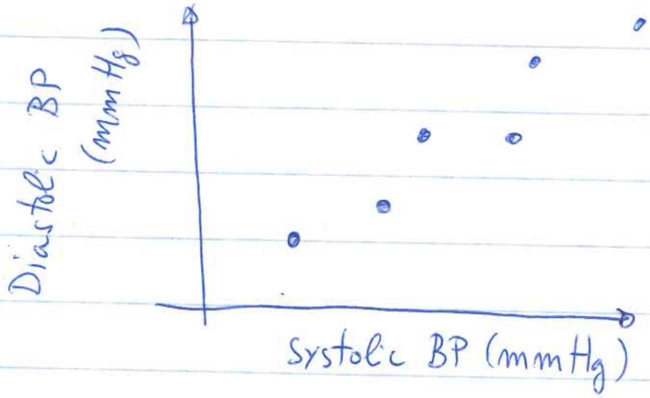
# PCA



## Eigen-decomposition of the covariance matrix

Example data




| Systolic BP | Diastolic BP |
|---|---|
| 126 | 78 |
| 128 | 80 |
| 128 | 82 |
| 130 | 82 |
| 130 | 84 |
| 132 | 86 |

We will use PCA to combine the two blood pressure variables into just one variable based on data from 6 individuals.

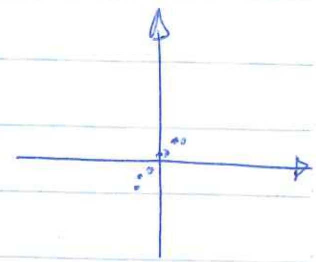
1. Center the data
2. Calculate the covariance matrix (CM)
3. Calculate eigenvalues of the CM
4. " eigenvectors of the CM
5. Order the eigenvectors
6. Calculate the principal components (PCs)

Note that sometimes we have to standardize the data by $\frac{x-\mu}{\sigma}$

Step 1:

| SBP | DBP |
|---|---|
| $126-129=-3$ | $78-82=-4$ |
| $128-129=-1$ | $80-82=-2$ |
| $128-129=-1$ | $82-82=0$ |
| $130-129=1$ | $82-82=0$ |
| $130-129=1$ | $84-82=2$ |
| $132-129=3$ | $86-82=4$ |

$\Rightarrow$



data centered around (0,0).

SBP $\Rightarrow$ cSBP

DBP $\Rightarrow$ cDBP

Step 2: Calculate the CM

$$\hat{\sigma} = \begin{array}{c|cc} & cSBP & cDBP \\ \hline cSBP & 4.4 & 5.6 \\ cDBP & 5.6 & 8.0 \end{array}$$

↳ variances (spread in cDBP is higher than in cSBP)

covariances

$$Var(cSBP) = \frac{1}{m-1} \sum_{i=1}^{m} \left( cSBP_i - \overline{cSBP} \right)^2$$

$$Var(cDBP) = \frac{1}{m-1} \sum_{i=1}^{m} \left( cDBP_i - \overline{cDBP} \right)^2$$

$$\sigma_{xy} = \frac{1}{m-1} \sum_{i=1}^{m} \left( cSBP_i - \overline{cSBP} \right)\left( cDBP_i - \overline{cDBP} \right)$$

Step 3. $\det \left| \hat{\sigma} - \lambda \hat{I} \right| = 0$

$$\det \begin{vmatrix} 4.4-\lambda & 5.6 \\ 5.6 & 8-\lambda \end{vmatrix} = 0 \quad \Rightarrow \quad 3.84 - 12.4\lambda + \lambda^2 = 0$$

$$\boxed{\begin{array}{l} \lambda_1 = 0.32 \\ \lambda_2 = 12.08 \end{array}}$$

↳ eigenvalues of the CM.

(2)

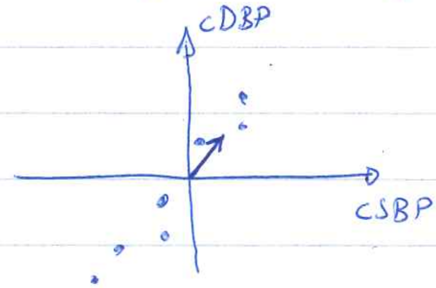
Step 4. $\hat{\sigma}\, v = \lambda\, v$ ↳ eigenvectors!

$$\begin{bmatrix} 4.4 & 5.6 \\ 5.6 & 8.0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \overset{\lambda_2}{\overset{\downarrow}{12.08}} \begin{bmatrix} x \\ y \end{bmatrix}$$

$\Rightarrow$ $5.6\,y = 7.68\,x$ solving for $y \Rightarrow y = 1.37\,x$

$5.6\,x = 4.08\,y$ for $x = 1 \Rightarrow y = 1.37$

Therefore: $v_2 = \begin{bmatrix} 1 \\ 1.37 \end{bmatrix}$




After normalization: $\boxed{v_2 = \begin{bmatrix} 0.59 \\ 0.81 \end{bmatrix}}$

For the other eigenvalue $\boxed{\lambda_1 = 0.32}$, we get $\boxed{v_1 = \begin{bmatrix} -0.81 \\ 0.59 \end{bmatrix}}$

(normalized)

Since the CM is a symmetric matrix, the eigenvectors will be <u>ORTHOGONAL</u> :

Step 5. Ordering eigen vector

The eigen vector with the largest eigen value becomes our first eigenvector.

$\nu_2 \rightarrow V_{PC1}$

$\nu_1 \rightarrow V_{PC2}$

We order these eigenvectors in a matrix called $\hat{V}$:

$$\hat{V} = \begin{array}{cc} V_{PC1} & V_{PC2} \\ \left[\begin{matrix} 0.59 \\ 0.81 \end{matrix}\right. & \left.\begin{matrix} -0.81 \\ 0.59 \end{matrix}\right] \end{array}$$

$\Uparrow$

Principal components

define a matrix $\hat{D}$ that has our centered data:

$$\hat{D} = \begin{bmatrix} -3 & -4 \\ -1 & -2 \\ -1 & 0 \\ 1 & 0 \\ 1 & 2 \\ 3 & 4 \end{bmatrix}$$

and make

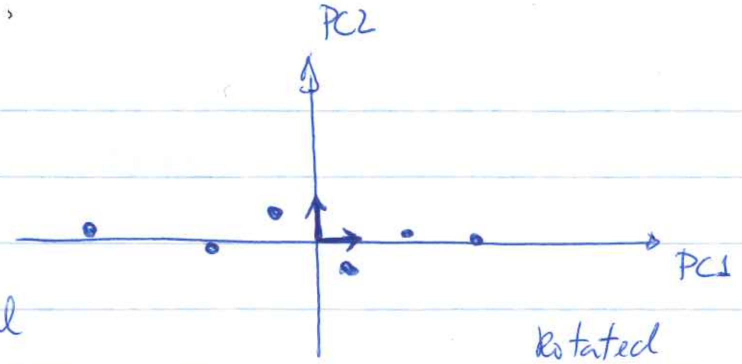
$$\hat{D}\hat{V} = \underbrace{\begin{array}{cc} PC1 & PC2 \\ \left[\begin{matrix} -5.0 \\ -2.2 \\ -0.6 \\ 0.6 \\ 2.2 \\ 5.0 \end{matrix}\right. & \left.\begin{matrix} 0.1 \\ -0.4 \\ 0.8 \\ -0.8 \\ 0.4 \\ -0.1 \end{matrix}\right] \end{array}}$$

transformed data! $\Rightarrow$ Principal component scores!

This represents the original centered data in the PC space!

# ROTATION






rotated plot also named score plot!

original

| cSBP | cDBP |
|---|---|
| −3 | −4 |
| −1 | −2 |
| −1 | 0 |
| 1 | 0 |
| 1 | 2 |
| 3 | 4 |
| Var = 4.4 | Var = 8.0 |

Rotated

| PC1 | PC2 |
|---|---|
| −5 | 0.1 |
| −2.2 | −0.4 |
| −0.6 | 0.8 |
| 0.6 | −0.8 |
| 2.2 | 0.4 |
| 5.0 | −0.1 |
| Var = 12.08 | Var = 0.32 |

eigen values

and variance [redacted] difference increased!

$$\% \text{Var} = \frac{\lambda_{PC1}}{\lambda_{PC1} + \lambda_{PC2}} = \frac{12.08}{12.08 + 0.32} = 97.4\%$$

PC1 captures 97.4 % of the total variance of the data!

$$\hat{\Sigma}_{PC} = \begin{pmatrix} 12.08 & 0 \\ 0 & 0.32 \end{pmatrix}$$

Rotation $\quad PC1 = 0.59\, cSBP + 0.81\, cDBP$

Transformed data $\quad PC2 = -0.81\, cSBP + 0.59\, cDBP$

centered data

For example: PC score for person #6 : $PC1_6 = 0.59 \times \boxed{3} + 0.81 \times \boxed{4} = 5$

But how variable reduction play a role in PCA?
We still have the same # of variables as of PCs.

Since the 1º PC captures ≈97% of all variance (carries most of the information about the data), we can <u>neglect PC2</u>.

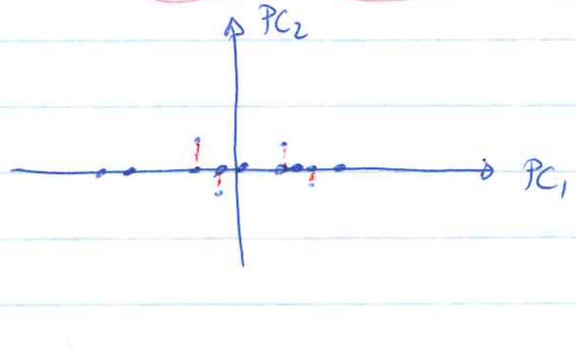
$$\boxed{PC1 = 0.59\, cSBP + 0.81\, cDBP}$$

We are combining the two variables cSBP and cDBP into 1 variable, the PC1, in a way that maximize the variance of the linear combination.

The weights tell how much each variable contributes to the PC.

$w_{cDBP} > w_{cSBP}$ : PCA puts more weight into cDBP when the 2 variables are combined.

NOTE: The covariance matrix transforms (rotates) any vector into the direction of the eigenvector of largest eigenvalue or variance!




ignoring PC2, data is projected into PC1.