From Statistics to Data Mining

Master 1
COlour in Science and Industry (COSI)
Cyber-Physical Social System (CPS2)
Saint-Étienne, France

Fabrice MUHLENBACH
https://perso.univ-st-etienne.fr/muhlfabr/
e-mail: fabrice.muhlenbach@univ-st-etienne.fr

Tutorial

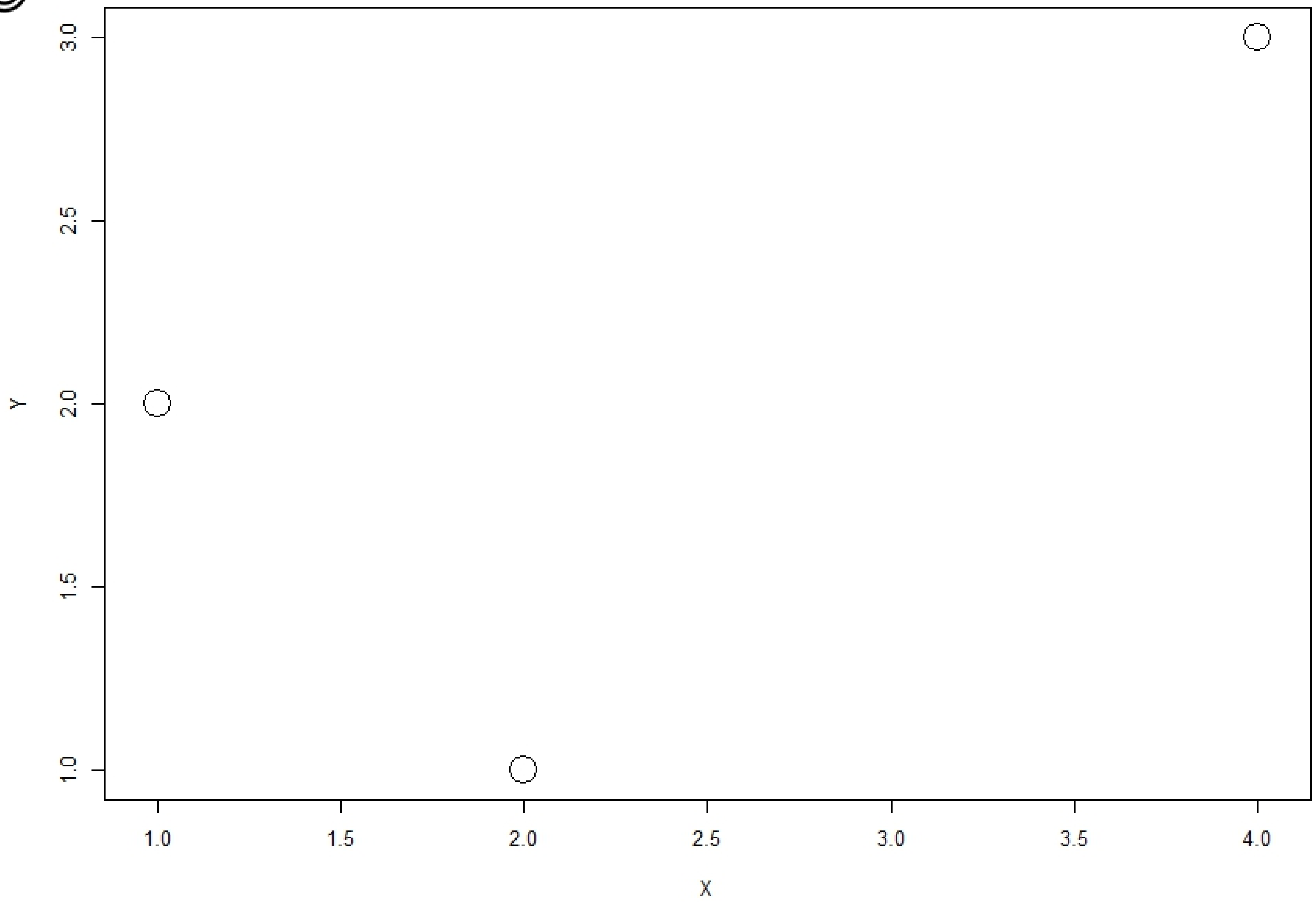
- Linear Regression

➢ let the following points:

x_i	1	2	4
y_i	2	1	3

➢ represent the points graphically

➢ then find the linear regression line
= the line fitting at best the points



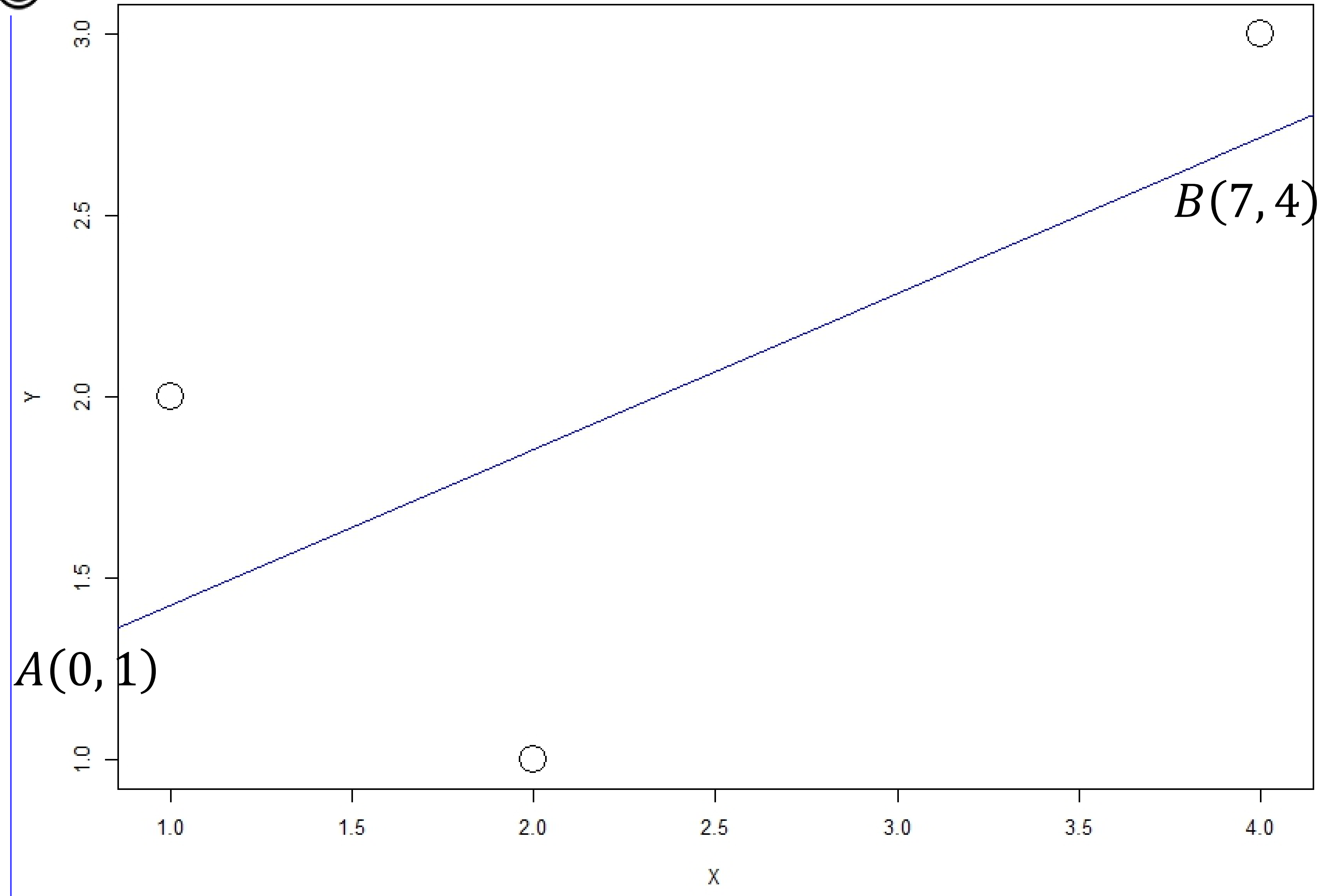
Y
X
1.0
1.5
2.0
2.5
3.0
3.5
4.0

- Linear Regression

- $\mathbb{R}^2$ → the parameters θ_0 and θ_1 are given by:

$$\theta_1 = \frac{\text{cov}(x,y)}{V(x)},\ \theta_0 = \bar{y} - \theta_1\bar{x}$$

- $\rightarrow y = \frac{\text{cov}(x,y)}{V(x)}x + \bar{y} - \frac{\text{cov}(x,y)}{V(x)}\bar{x}$
- $\bar{x} = \frac{1+2+4}{3} = \frac{7}{3} \qquad \bar{y} = \frac{2+1+3}{3} = 2$
- $\overline{x.y} = \frac{1\times2+2\times1+4\times3}{3} = \frac{16}{3} \rightarrow cov(x,y) = \overline{x.y} - \bar{x}.\bar{y} = \frac{16-7\times2}{3} = \frac{2}{3}$
- $\overline{x^2} = \frac{1^2+2^2+4^2}{3} = 7 \rightarrow V(x) = \overline{x^2} - \bar{x}^2 = 7 - \left(\frac{7}{3}\right)^2 \rightarrow \theta_1 = \frac{3}{7}$
- $\theta_0 = \bar{y} - \theta_1\bar{x} = 2 - \frac{3}{7}\times\frac{7}{3} = 2 - 1 = 1 \rightarrow y = \frac{3}{7}x + 1$


B(7, 4)
A(0, 1)
Y
3.0
2.5
2.0
1.5
1.0
1.0
1.5
2.0
2.5
3.0
3.5
4.0
X

Tutorial

- Principal Component Analysis

➢ let the following points:

x_i	1	3	3	5	5	6	8	9
y_i	2	3	5	4	6	5	7	8

➢ plot the data in the original 2D-space

➢ compute the covariance matrix Σ from the zero mean values $(x - \bar{x})$

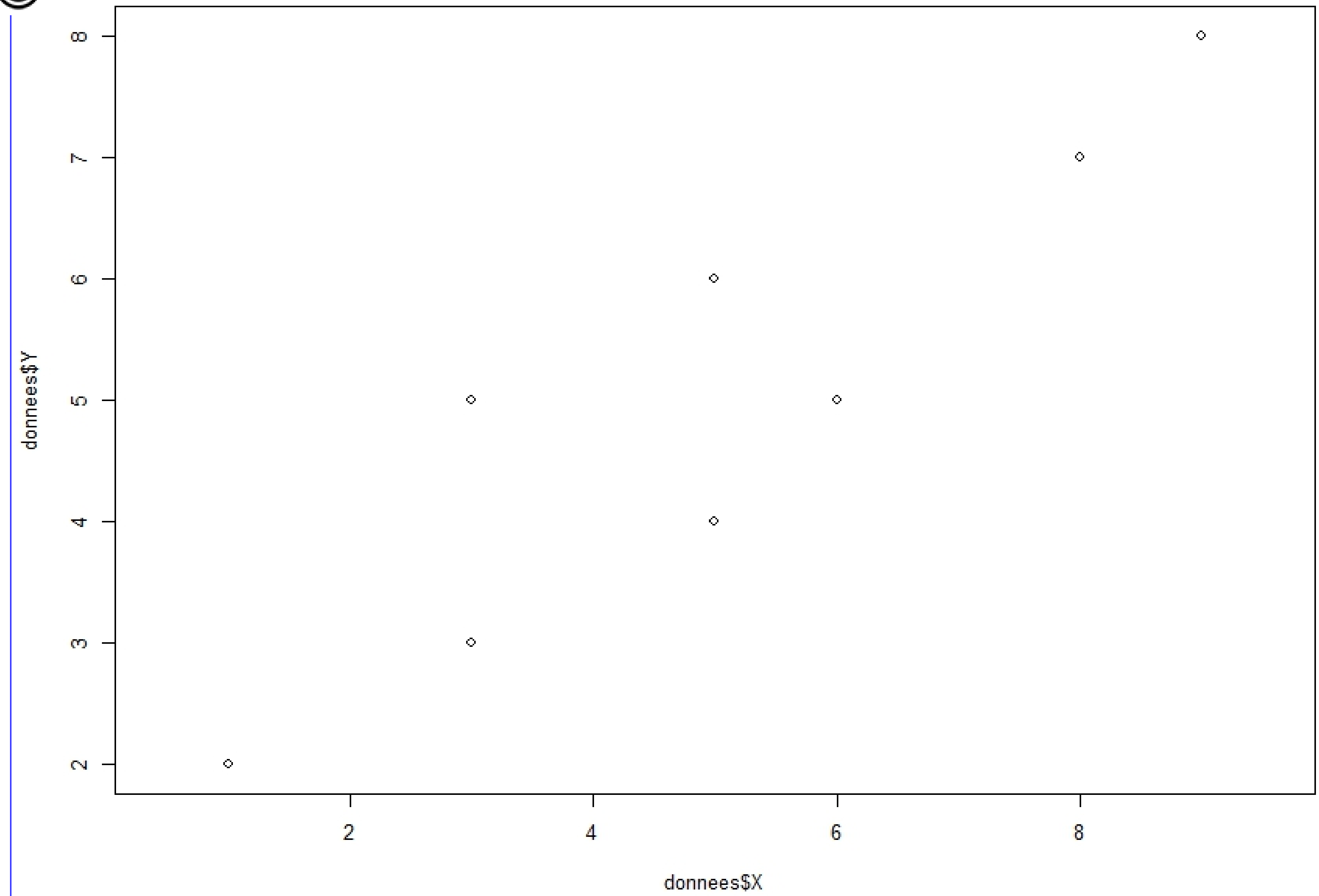
➢ solve the characteristic equation $det(\Sigma - \lambda I) = 0$ to get the eigenvalues

➢ deduce the first eigenvector $\mathbf{u}_1$ from the largest eigenvalue λ_1 by solving $\mathbf{u}_1\Sigma = \lambda_1\mathbf{u}_1$ (nota: first assume that one of the variables is equal to 1, then find the other one and finally normalize the vector to make it unit-length)

➢ use $\mathbf{u}_1$ to find the projection $t_i = \mathbf{u}_1{}^{\mathrm{T}} x_i$ for every training data x_i

➢ plot the points in the new space according to the first component $\mathbf{u}_1$

➢ compute the part of the total variance explained by this 1-D space.



donnees$Y
donnees$X
2
3
4
5
6
7
8
2
4
6
8

Tutorial

• Principal Component Analysis

➢ compute the covariance matrix Σ from the zero mean values $(x - \bar{x})$:

○ covariance matrix $\Sigma = \begin{pmatrix} \mathrm{cov}(X,X) & \mathrm{cov}(X,Y) \\ \mathrm{cov}(Y,X) & \mathrm{cov}(Y,Y) \end{pmatrix}$

$$\bar{x} = \frac{1+3+3+5+5+6+8+9}{8} = \frac{40}{8} = 5$$

$$\bar{y} = \frac{2+3+5+4+6+5+7+8}{8} = \frac{40}{8} = 5$$

x_i	1	3	3	5	5	6	8	9	$\Sigma = 40$
y_i	2	3	5	4	6	5	7	8	$\Sigma = 40$
$x_i - \bar{x}$	-4	-2	-2	0	0	1	3	4	$\Sigma = 0$
$y_i - \bar{y}$	-3	-2	0	-1	1	0	2	3	$\Sigma = 0$
$(x_i - \bar{x})^2$	16	4	4	0	0	1	9	16	$\Sigma = 50$
$(y_i - \bar{y})^2$	9	4	0	1	1	0	4	9	$\Sigma = 28$
$(x_i - \bar{x}) \times (y_i - \bar{y})$	12	4	0	0	0	0	6	12	$\Sigma = 34$

Tutorial

- Principal Component Analysis

➢ compute the covariance matrix Σ from the zero mean values $(x - \bar{x})$:

○ covariance matrix $\Sigma = \begin{pmatrix} \mathrm{cov}(X,X) & \mathrm{cov}(X,Y) \\ \mathrm{cov}(Y,X) & \mathrm{cov}(Y,Y) \end{pmatrix}$

○ $\mathrm{cov}(X,X) = \frac{50}{8} = \frac{25}{4} = 6.25$

○ $\mathrm{cov}(X,Y) = \mathrm{cov}(Y,X) = \frac{34}{8} = \frac{17}{4} = 4.25$

○ $\mathrm{cov}(Y,Y) = \frac{28}{8} = \frac{7}{2} = 3.5$

○ → covariance matrix: $\Sigma = \begin{pmatrix} \mathrm{cov}(X,X) & \mathrm{cov}(X,Y) \\ \mathrm{cov}(Y,X) & \mathrm{cov}(Y,Y) \end{pmatrix} = \begin{pmatrix} \frac{25}{4} & \frac{17}{4} \\ \frac{17}{4} & \frac{7}{2} \end{pmatrix}$

Tutorial

• Principal Component Analysis

- ➢ solve the characteristic equation $det(\Sigma - \lambda I) = 0$ to get the eigenvalues
- ○ $\det(\Sigma - \lambda I) = 0$ $\Leftrightarrow \det\begin{pmatrix} \frac{25}{4} - \lambda & \frac{17}{4} \\ \frac{17}{4} & \frac{7}{2} - \lambda \end{pmatrix} = 0$

$$\Leftrightarrow \det\begin{pmatrix} \frac{25-4\lambda}{4} & \frac{17}{4} \\ \frac{17}{4} & \frac{14-4\lambda}{4} \end{pmatrix} = 0$$

$$\Leftrightarrow \frac{25-4\lambda}{4} \times \frac{14-4\lambda}{4} - \frac{17}{4} \times \frac{17}{4} = 0$$

$$\Leftrightarrow \frac{(25-4\lambda) \times (14-4\lambda) - 17^2}{4^2} = 0$$

$$\Leftrightarrow (25-4\lambda) \times (14-4\lambda) - 17^2 = 0$$

$$\Leftrightarrow 350 - 100\lambda - 56\lambda + 16\lambda^2 - 289 = 0$$

$$\Leftrightarrow 16\lambda^2 - 156\lambda + 61 = 0$$

Tutorial

• Principal Component Analysis

- ➢ solve the characteristic equation $det(\Sigma - \lambda I) = 0$ to get the eigenvalues
- ○ $\det(\Sigma - \lambda I) = 0 \Leftrightarrow 16\lambda^2 - 156\lambda + 61 = 0$ → discriminant :

$$\delta = (-156)^2 - 4 \times 16 \times 61 = 20432$$

- ○ solutions are $\lambda_1 = \frac{156+\sqrt{\delta}}{2\times16} = 9.34$ and $\lambda_2 = \frac{156-\sqrt{\delta}}{2\times16} = 0.41$
- ➢ deduce the first eigenvector $\mathbf{u}_1$ from the largest eigenvalue λ_1 by solving $\mathbf{u}_1\Sigma = \lambda_1\mathbf{u}_1 \Rightarrow \Sigma e_i = \lambda e_i$

$$\Leftrightarrow \begin{pmatrix} \frac{25}{4} & \frac{17}{4} \\ \frac{17}{4} & \frac{14}{4} \end{pmatrix} \begin{pmatrix} e_{1,1} \\ e_{1,2} \end{pmatrix} = \lambda_1 \begin{pmatrix} e_{1,1} \\ e_{1,2} \end{pmatrix}$$

$$\Leftrightarrow \left\{ \begin{array}{rcl} \frac{25}{4}e_{1,1} + \frac{17}{4}e_{1,2} & = & \lambda_1 e_{1,1} \\ \frac{17}{4}e_{1,1} + \frac{14}{4}e_{1,2} & = & \lambda_1 e_{1,2} \end{array} \right\}$$

$$\Leftrightarrow \left(\frac{25 - 4\lambda_1}{4}\right) e_{1,1} = -\frac{17}{4} e_{1,2}$$

Tutorial

• Principal Component Analysis

➤ deduce the first eigenvector $\mathbf{u}_1$ from the largest eigenvalue λ_1 by solving $\mathbf{u}_1\Sigma = \lambda_1\mathbf{u}_1 \Rightarrow \Sigma e_i = \lambda e_i$

$$\Leftrightarrow \left(\frac{25-4\lambda_1}{4}\right) e_{1,1} = -\frac{17}{4} e_{1,2}$$

$$\Leftrightarrow e_{1,1} = -\frac{17}{25-4\lambda_1} e_{1,2} = 1,374563 \times e_{1,2}$$

$$\Leftrightarrow e_1 \sim \begin{bmatrix} 1,374563 \\ 1 \end{bmatrix}$$

$$\|e_1\| = \sqrt{(1,374563)^2 + 1^2} = 1,69983.$$

○ we divide the eigenvector e_1 by its norm to have the unit vector $\mathbf{u}_1$:

$$u_1 \sim \begin{bmatrix} 0,8086471 \\ 0,588294 \end{bmatrix} \quad e_{2,1} = -\frac{17}{25-4\times\lambda_2} e_{2,2}, \quad e_2 \sim \begin{bmatrix} -0.727504 \\ 1 \end{bmatrix}$$

$$u_2 \sim \begin{bmatrix} -0,588294 \\ 0,8086471 \end{bmatrix}$$

Tutorial

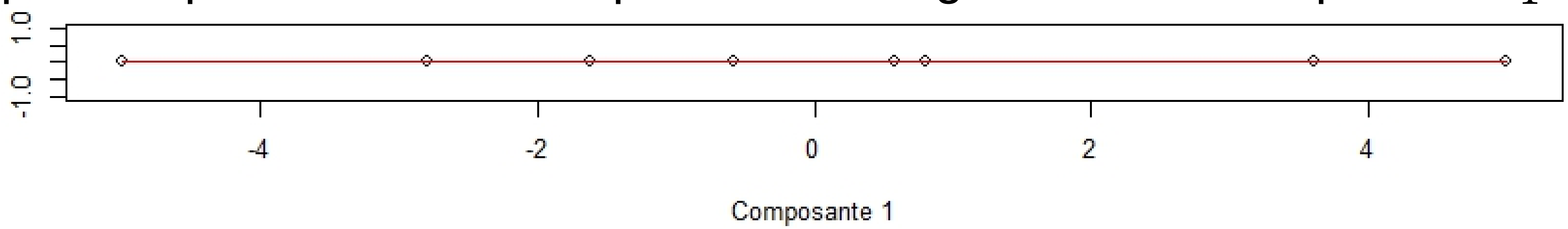
• Principal Component Analysis

- use $\mathbf{u}_1$ to find the projection $t_i = \mathbf{u}_1{}^{\mathrm{T}} x_i$ for every training data x_i

$$c_1 = (x_i - \bar{x}) \times u_{1,1} + (y_i - \bar{y}) \times u_{1,2}.$$
$$c_2 = (x_i - \bar{x}) \times u_{2,1} + (y_i - \bar{y}) \times u_{2,2}.$$

x_i	1	3	3	5	5	6	8	9
y_i	2	3	5	4	6	5	7	8
c_1	-4.9994705	-2.7938823	-1.6172942	-0.5882940	0.5882940	0.8086471	3.6025294	4.9994705
c_2	-0.07276523	-0.44070617	1.17658805	-0.80864711	0.80864711	-0.58829402	-0.14758786	0.07276523

- plot the points in the new space according to the first component $\mathbf{u}_1$



- compute the part of the total variance explained by this 1-D space:

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = 0.96$$