PCA Homework

PCA on the Given Data Set

Centering the Data

Given the data:

Instance	x	y
m_1	0.2	-0.3
m_2	-1.1	2
m_3	1	-2.2
m_4	0.5	-1
m_5	-0.6	1
Mean	0	-0.1

Center the data by subtracting the mean:

$$x_i' = x_i - \bar{x} = x_i - 0 = x_i$$
$$y_i' = y_i - \bar{y} = y_i - (-0.1) = y_i + 0.1$$

Centered data:

Instance	x_i'	y_i'
m_1	0.2	-0.2
m_2	-1.1	2.1
m_3	1	-2.1
m_4	0.5	-0.9
m_5	-0.6	1.1

Computing the Covariance Matrix

Calculate the variances and covariance:

$$\sigma_x^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i')^2$$
$$\sigma_y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(y_i')^2$$
$$\sigma_{xy} = \frac{1}{m-1}\sum_{i=1}^{m}x_i'y_i'$$

Compute the sums:

$$\sum(x_i')^2 = 0.2^2 + (-1.1)^2 + 1^2 + 0.5^2 + (-0.6)^2 = 2.86$$
$$\sum(y_i')^2 = (-0.2)^2 + 2.1^2 + (-2.1)^2 + (-0.9)^2 + 1.1^2 = 10.88$$
$$\sum x_i'y_i' = (0.2)(-0.2) + (-1.1)(2.1) + (1)(-2.1) + (0.5)(-0.9) + (-0.6)(1.1) = -5.56$$

Compute variances and covariance:

$$\sigma_x^2 = \frac{2.86}{4} = 0.715$$
$$\sigma_y^2 = \frac{10.88}{4} = 2.72$$
$$\sigma_{xy} = \frac{-5.56}{4} = -1.39$$

Covariance matrix C:

$$C = \begin{bmatrix} 0.715 & -1.39 \\ -1.39 & 2.72 \end{bmatrix}$$

Calculating Eigenvalues and Eigenvectors

Solve for eigenvalues λ:

$$\det(C - \lambda I) = 0$$

Compute the determinant:

$$\begin{aligned} \det(C - \lambda I) &= (0.715 - \lambda)(2.72 - \lambda) - (-1.39)^2 \\ &= (0.715 - \lambda)(2.72 - \lambda) - 1.9321 \\ &= [1.9448 - 3.435\lambda + \lambda^2] - 1.9321 \\ &= \lambda^2 - 3.435\lambda + 0.0127 = 0 \end{aligned}$$

Solve the quadratic equation:

$$\lambda = \frac{3.435 \pm \sqrt{(3.435)^2 - 4 \times 0.0127}}{2} \approx \begin{cases} \lambda_1 \approx 3.432 \\ \lambda_2 \approx 0.003 \end{cases}$$

Eigenvalues:

$$\lambda_1 \approx 3.432, \quad \lambda_2 \approx 0.003$$

Compute eigenvectors.

For λ_1:

$$(C - \lambda_1 I)\mathbf{v} = \mathbf{0}$$

Compute:

$$\begin{bmatrix} 0.715 - 3.432 & -1.39 \\ -1.39 & 2.72 - 3.432 \end{bmatrix} = \begin{bmatrix} -2.717 & -1.39 \\ -1.39 & -0.712 \end{bmatrix}$$

Set up equations:

$$\begin{cases} -2.717v_1 - 1.39v_2 = 0 \\ -1.39v_1 - 0.712v_2 = 0 \end{cases}$$

Solve for v_2:

$$v_2 = \left(\frac{-2.717}{1.39}\right)v_1 \approx (-1.955)v_1$$

Normalized eigenvector:

$$\mathbf{v}_1 = \frac{1}{\sqrt{1 + (-1.955)^2}} \begin{bmatrix} 1 \\ -1.955 \end{bmatrix} \approx \begin{bmatrix} 0.4555 \\ -0.8902 \end{bmatrix}$$

For λ_2:

Similarly, we find:

$$\mathbf{v}_2 \approx \begin{bmatrix} 0.8902 \\ 0.4555 \end{bmatrix}$$

Transforming the Data

Transformation matrix W (using the first eigenvector):

$$W = \begin{bmatrix} 0.4555 \\ -0.8902 \end{bmatrix}$$

Compute the transformed data $\mathbf{Z}$:

$$Z_i = x_i' \times 0.4555 + y_i' \times (-0.8902)$$

Compute for each instance:

$$
\begin{aligned}
Z_1 &= 0.4555 \times 0.2 - 0.8902 \times (-0.2) = 0.2691 \\
Z_2 &= 0.4555 \times (-1.1) - 0.8902 \times (2.1) = -2.3703 \\
Z_3 &= 0.4555 \times 1 - 0.8902 \times (-2.1) = 2.3249 \\
Z_4 &= 0.4555 \times 0.5 - 0.8902 \times (-0.9) = 1.0290 \\
Z_5 &= 0.4555 \times (-0.6) - 0.8902 \times (1.1) = -1.2525
\end{aligned}
$$

Transformed data:

Instance	Z_i
m_1	0.2691
m_2	-2.3703
m_3	2.3249
m_4	1.0290
m_5	-1.2525

Percentage of Total Information in the First Principal Component

Total variance:

$$\text{Total Variance} = \lambda_1 + \lambda_2 = 3.432 + 0.003 = 3.435$$

Percentage of variance explained by the first principal component:

$$\text{Percentage} = \left(\frac{\lambda_1}{\text{Total Variance}}\right) \times 100\% = \left(\frac{3.432}{3.435}\right) \times 100\% \approx 99.91\%$$

Answer: The first principal component retains approximately 99.91% of the total information.