Prof. Dr. Jens Honore Walther
Dr. Georgios Arampatzis
ETH Zentrum, CLT
CH-8092 Zürich

Solution Set 11

Issued: 28.05.2021

Question 1: Dimensionality Reduction with PCA

a) The mean is simply given by

$$\bar{\mathbf{x}} = \frac{1}{8}\begin{pmatrix}24 & 24\end{pmatrix} = \begin{pmatrix}3 & 3\end{pmatrix}. \tag{1}$$

The centered data-matrix is thus:

$$X_{cent} = X - \bar{\mathbf{x}} = \begin{pmatrix}-2 & -2\\ -1 & -2\\ -1 & 0\\ 0 & -1\\ 0 & 1\\ 1 & 0\\ 1 & 1\\ 2 & 3\end{pmatrix}. \tag{2}$$

The data covariance matrix is

$$\hat{C} = \frac{1}{N-1}X_{cent}^\top X_{cent} = \frac{1}{7}\begin{pmatrix}4+1+1+1+1+4 & 4+2+1+6\\ 4+2+1+6 & 4+4+1+1+1+9\end{pmatrix} \implies$$

$$\hat{C} = \frac{1}{7}\begin{pmatrix}12 & 13\\ 13 & 20\end{pmatrix} \approx \begin{pmatrix}1.71 & 1.86\\ 1.86 & 2.86\end{pmatrix}$$

b) We know that $\hat{C}\mathbf{v}_1 = \lambda_1\mathbf{v}_1$, where $\mathbf{v}_1$ is the principal eigenvector. Assume that $\mathbf{v}_1 = (v_x, v_y)$, we have that

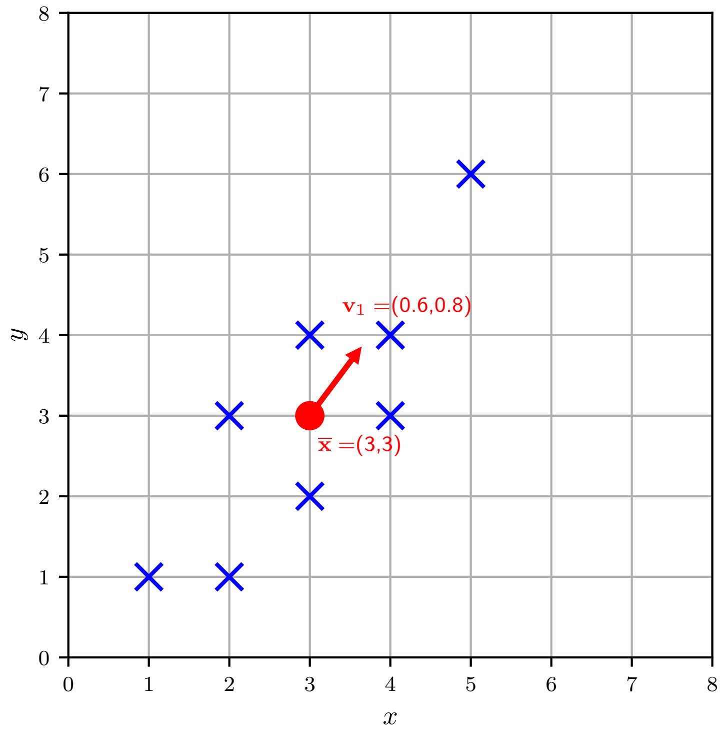
$$1.71\,v_x + 1.86\,v_y = 4.23\,v_x \implies 1.34\,v_x = v_y \quad \Big(\implies v_x = 0.74\,v_y\Big) \tag{3}$$

Moreover, we know that $\|\mathbf{v}_1\|_2 = 1$. As a result

$$v_x^2 + v_y^2 = 1 \implies v_x^2(1+1.34^2) = 1 \implies 2.8\,v_x^2 = 1 \implies \tag{4}$$

$$v_x = \sqrt{1/2.8} \approx 0.6. \tag{5}$$

and $v_y = \sqrt{1-v_x^2} = 0.8$. The leading eigenvector becomes $\mathbf{v}_1 = (0.6, 0.8)$, and is plotted below:



c) The eigenvalues give a measure of the variance of the distribution of X on each projection. Projecting on the first eigenvector of the covariance matrix will lead to a reduced order space retaining:

$$\frac{\lambda_1}{\sum_{i=1}^{2} \lambda_i} = \frac{4.23}{4.23 + 0.34} = 0.925 = 92.5\%, \tag{6}$$

of the total data variance.

d) The variance in the original two-dimensional space is

$$Var[X] = Var[X_{cent}] = \frac{1}{7}\begin{pmatrix} 4+1+1+1+1+4 \\ 4+4+1+1+1+9 \end{pmatrix} = \frac{1}{7}\begin{pmatrix} 12 \\ 20 \end{pmatrix} \approx \begin{pmatrix} 1.71 \\ 2.86 \end{pmatrix}. \tag{7}$$

which can also be obtained considering the definition of the covariance matrix, as it contains the variances in its diagonal:

$$\sigma^2 = \mathbf{u}^\top C \mathbf{u} \tag{8}$$

As a consequence, the total variance in the original space is $\sigma^2 = \sigma_x^2 + \sigma_y^2 \approx 4.57$.
Using the principal component $V_r = \mathbf{v}_1^T \in \mathbb{R}^{2\times 1}$, we can project the data on a one-dimensional manifold. The projected (centered) data is given by

$$Z = X_{cent}V_r = X_{cent}\mathbf{v}_1^T = \begin{pmatrix} -2 & -2 \\ -1 & -2 \\ -1 & 0 \\ 0 & -1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \\ 2 & 3 \end{pmatrix} \begin{pmatrix} 0.6 \\ 0.8 \end{pmatrix} = \begin{pmatrix} -2.8 \\ -2.2 \\ -0.6 \\ -0.8 \\ 0.8 \\ 0.6 \\ 1.4 \\ 3.6 \end{pmatrix}, \tag{9}$$

with mean $\overline{Z} = 0$.

The total explained variance in the reduced order subspace is:

$$Var[Z] = \mathbb{E}[(Z - \overline{Z})^2] \tag{10}$$

The variance is given by

$$\sigma_Z^2 = \frac{2.8^2 + 2.2^2 + 2 \times 0.6^2 + 2 \times 0.8^2 + 1.4^2 + 3.6^2}{7} = \frac{29.6}{7} = 4.23 \tag{11}$$

which corresponds to $4.23/4.57 = 92.5\%$ of the total variance, which corresponds to the value obtained using the eigenvalue approach.

e) PCA can be used as a dimensionality reduction technique to project the data to a reduced order space, capturing the data's variance as much as possible. This way, we can first apply PCA to the data as a pre-processing step, and then use the desired algorithm in the reduced space, which will make computations hopefully tractable.