ETH Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

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

Solution Set 13

Issued: 29.05.2020

In this exercise, you will learn how to do dimensionality reduction using Principle Component Analysis (PCA). The accompanying notebook will allow you to apply this knowledge in a practical example.

Question 1: Dimensionality Reduction with PCA

In the following we want to see how we can use PCA to perform dimensionality reduction. In the following we assume that you are given a dataset of measurements $\mathbf{x}_i \in \mathbb{R}^D$ for $i = 1, \ldots, N$. In a first step we want to transform the data such that is has zero mean (i.e. $\sum \tilde{\mathbf{x}}_i = 0$ for $\tilde{\mathbf{x}}_i = \mathbf{x}_i - \bar{x}$, where $\bar{x} = \frac{1}{N} \sum \mathbf{x}_i$) and combine it in a matrix $\tilde{X} = (\tilde{\mathbf{x}}_1, \ldots, \tilde{\mathbf{x}}_n) \in \mathbb{R}^{D \times N}$. We have seen in the notes that the covariance matrix $C \in \mathbb{R}$ of this data can be written as

$$C = \frac{1}{N-1} \tilde{X}^\top \tilde{X} \tag{1}$$

In the lecture you have seen, that finding the direction in the data which maximizes is obtained by performing an eigenvalue decomposition

$$C \mathbf{v}_i = \lambda_i \mathbf{v}_i \quad \text{for} \quad i = 1, \ldots, N \tag{2}$$

Where the directions that maximize the variance are given by sorting the eigenvectors with respect of the size of their eigenvalues $\lambda_{(1)} < \cdots < \lambda_{(N)}$. The first principal component is then $\mathbf{v}_{(1)}$, the second $\mathbf{v}_{(2)}$ and so on. They describe the variance in the data in descending order.They describe the variance in the data in descending order. As also remarked in the lecture, this can become computationally very heavy if the number of data points becomes very large $N \gg 1$. A way around is to use

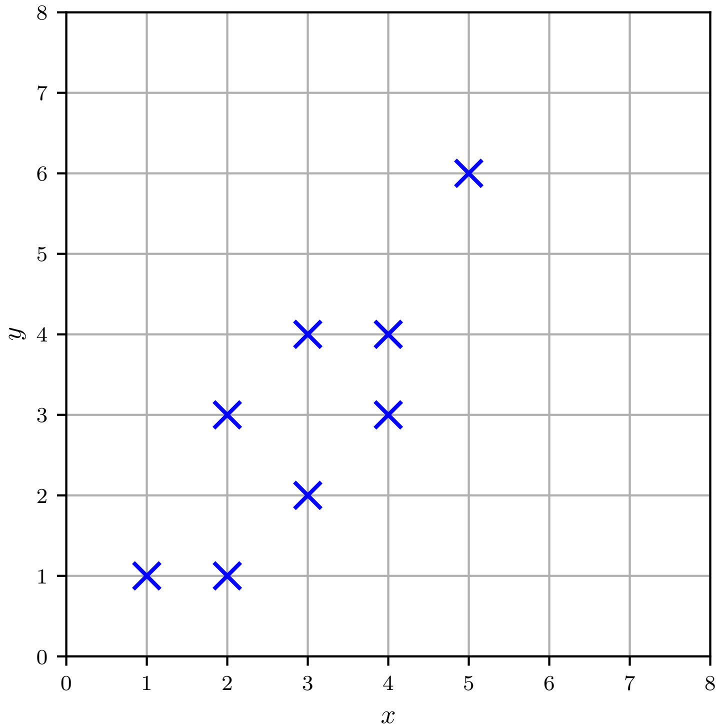
$$\hat{C} = \frac{1}{N-1} \tilde{X} \tilde{X}^\top \tag{3}$$

instead. We see that $\hat{C} \in \mathbb{R}^{D \times D}$ and thus the problem of solving a N-dimensional eigenproblem is reduced so solve a D-dimensional eigenproblem.

In the following, you are given a collection of $N = 8$ data points in a two dimensional space:

$$X^\top = \begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 2 & 3 \\ 3 & 2 \\ 3 & 4 \\ 4 & 3 \\ 4 & 4 \\ 5 & 6 \end{pmatrix}, \tag{4}$$

with $X^\top \in \mathbb{R}^{N\times D}$, plotted below:



a) The mean is given by

$$\bar{\mathbf{x}} = \frac{1}{8}\begin{pmatrix}24 & 24\end{pmatrix} = \begin{pmatrix}3 & 3\end{pmatrix}. \tag{5}$$

The centered data-matrix are thus:

$$\tilde{X} = X - \bar{\mathbf{x}} = \begin{pmatrix} -2 & -2 \\ -1 & -2 \\ -1 & 0 \\ 0 & -1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \\ 2 & 3 \end{pmatrix}. \tag{6}$$

The data covariance matrix is

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

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

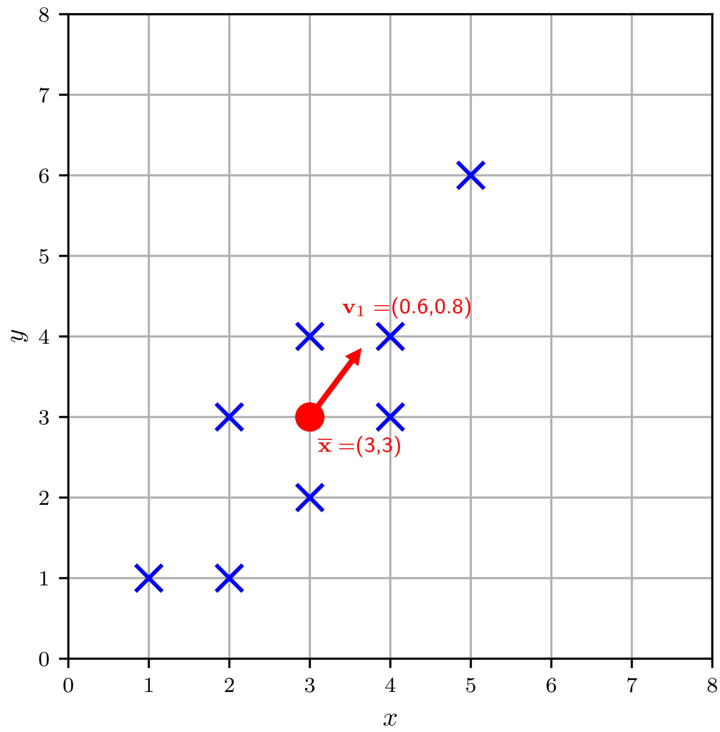
b) We know that $\hat{C}\mathbf{v}_1 = \lambda_1\mathbf{v}_1$, where $\mathbf{v}_1$ is the principal eigenvector. Assume that $\mathbf{v}_1 = (v_x, v_y)$, we have that

$$1.71\, v_x + 1.86\, v_y = 4.23\, v_x \implies 1.34\, v_x = v_y \quad \left(\implies v_x = 0.74\, v_y \right) \tag{7}$$

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

$$v_x^2 + v_y^2 = 1 \implies v_x^2(1 + 1.34^2) = 1 \implies 2.8\, v_x^2 = 1 \implies \tag{8}$$
$$v_x = \sqrt{1/2.8} \approx 0.6. \tag{9}$$

As a consequence, $v_y = \sqrt{1 - v_x^2} = 0.8$. The leading eigenvector is thus $\mathbf{v}_1 = (0.6, 0.8)$. The eigenvector 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:

$$\frac{\lambda_1}{\sum_{i=1}^{2} \lambda_i} = \frac{4.23}{4.23 + 0.34} = 0.925 = 92.5\%, \tag{10}$$

of the total data variance.

d) The variance $\sigma^2 = \mathbb{E}[x^2]$ in the original two dimensional space is

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{7}\left(\begin{pmatrix}-2\\-2\end{pmatrix}^2 + \begin{pmatrix}-1\\-2\end{pmatrix}^2 \begin{pmatrix}-1\\0\end{pmatrix}^2 \begin{pmatrix}0\\-1\end{pmatrix}^2 \begin{pmatrix}0\\1\end{pmatrix}^2 \begin{pmatrix}1\\0\end{pmatrix}^2 \begin{pmatrix}1\\1\end{pmatrix}^2 \begin{pmatrix}2\\3\end{pmatrix}^2\right) \\ &= \frac{1}{7}\left((4+4) + (1+4) + (1+0) + (0+1) + (1+0) + (1+1) + (4+9)\right) \\ &= 4.57 \end{aligned} \tag{11}$$

Thus, the total variance in the original space is $\sigma^2 = \sigma_x^2 + \sigma_y^2 \approx 4.57$.

Using the principal component $V_r = \mathbf{u}_1^T \in \mathbb{R}^{2\times 1}$, we can project the data on a one dimensional manifold. The projected (centered) data are given by

$$Z = \tilde{X}V_r = \tilde{X}\mathbf{u}_1^T = \begin{pmatrix} -2 & -2 \\ -1 & -2 \\ -1 & 0 \\ 0 & -1 \\ 0 & 1 \\ 1 & 0 \\ 1 & 1 \\ 2 & 3 \end{pmatrix} \begin{pmatrix} 0.6 \\ 0.8 \end{pmatrix} = \begin{pmatrix} -2.8 \\ -2.2 \\ -0.6 \\ -0.8 \\ 0.8 \\ 0.6 \\ 1.4 \\ 3.6 \end{pmatrix}, \tag{12}$$

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

The total explained variance in the reduced order subspace is:

$$\sigma^2(Z) = \mathbb{E}[(Z - \overline{Z})^2] \tag{13}$$

The variance is given by

$$\sigma_Z^2 = \frac{2.8^2 + 2.2^2 + 2 \times 0.6^2 + 2 \times 0.8^2 + 1.4^2 + 3.6^2}{7} = \frac{29.6}{7} = 4.23 \tag{14}$$

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

e) PCA can be used as a dimensionality reduction technique to project the data to a reduced order space, capturing as much data variance as possible. In this way, we first can apply PCA as a pre-processing step to the data and then use the desired algorithm on the reduced order tractable space.