Worksheet on "PCA"

CS5691 PRML Jul–Nov 2025

September 26, 2025

1. Consider a dataset of three data points in 2-D space: $(1, 1)$, $(2, 2)$, and $(3, 3)$.

   (a) What is the first principal component, denoted by the vector $PC1$, of this dataset?

   (b) If we want to project the original data points into 1-D space defined by "$PC1$", what is the variance of the projected data along this component $PC1$?

   (c) Let us use the projection onto $PC1$ to approximate each datapoint $x_n \in \mathbb{R}^2$ by another data point $\tilde{x}_n \in \mathbb{R}^2$ to minimize the total sum of squares erorr between the original and approximated datapoints. Then, what are the resulting approximate datapoints $\tilde{x}_n \in \mathbb{R}^2$, and what is the resulting reconstruction error?

2. Given a dataset $X \in \mathbb{R}^{N \times D}$ (i.e., N data points along the rows, and D features along the columns):

$$X = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

   Calculate the covariance matrix and the corresponding eigenvectors to find all the principal components of the dataset. Determine the minimum number of principal components required to retain at least 90% of the variance in the dataset.

3. Consider the following dataset D of 4 datapoints:

data #	x	y
1	3	2
2	4	2
3	5	3
4	6	3

Table 1: Dataset D

   You need to reduce the data into a single-dimension representation. You are given the first principal component: $PC1 \approx \begin{bmatrix} 0.92 \\ 0.38 \end{bmatrix}$.

   (a) (2 points) What is the xy coordinate for the datapoint reconstructed (approximated) from data #1 (x=3, y=2) using the first principal component of D? What is the reconstruction error of this $PC1$-based approximation of data #1?

   (b) (2 points) What is the second principal component of the dataset D? How will you represent data #1 as a linear combination of the two principal components? What is the reconstruction error of this $(PC1, PC2)$-based representation of data #1?

   (c) (2 points) Let D' be the mean-subtracted version of D. What will be the first and second principal components $PC1$ and $PC2$ of D'? What is the xy coordinate of data #1 and its $PC1$-based reconstruction in D'? What is the associated reconstruction/approximation error of data #1?