Assignment 4: CS 215

Devansh Jain	Harshit Varma
190100044	190100055

November 13, 2020

Question 3

Instructions for running the code:

1. Unzip and `cd` to `q3/code`, under this find the file named `q3.m`
2. Run the file, two scatter plots with the fitted lines for each of the datasets will generated and saved to the `q3/results` folder.

We have been given observed set of data $\{\mathbf{z}_i\}_{i=1}^N$, where $\mathbf{z}_i = (x_i, y_i)$
We need to find a linear relationship between the RVs X and Y

Thus, we need to find m, c such that $Y \approx mX + c$ using PCA.
This is essentially dimensionality reduction from $\mathbb{R}^2$ to $\mathbb{R}$
Thus, we need to find the first principal mode of variance/principal component of $\{\mathbf{z}_i\}_{i=1}^N$

Let the sample mean be $\mu = \left(\mu_1 = \frac{\sum_{i=1}^N x_i}{N},\ \mu_2 = \frac{\sum_{i=1}^N y_i}{N}\right)$, then first we center the data about μ.
Let the sample covariance matrix of the new centered data be C.
Since we are not allowed to use `cov()` and `mean()`, we compute the sample $\mu = \frac{\sum_{i=1}^N \mathbf{z_i}}{N}$.
Let $Z = [\mathbf{z_i} - \mu]$ be a $N \times 2$ matrix of the entire **mean-centered** observed data.
Then $C = \frac{Z^T Z}{N-1}$ will give us the covariance matrix.

On eigenvalue decomposition of C, we get $C = Q\Lambda Q^T$, let $\mathbf{d}$ correspond to the diagonal values of Λ, let $j = \text{argmax}_{i:1 \le i \le 2}\ \mathbf{d}$, then the j^{th} column of Q corresponds to the principal mode of variation, $\mathbf{v} = (v_1, v_2)$

Thus, $m = \frac{v_2}{v_1}$, and $y = mx$, since this was the centered coordinate system (about μ), we need to shift back to the original coordinate system, thus, the final straight line will be $y - \mu_2 = m(x - \mu_1)$

The same method as above has been implemented in `q3.m`

In this case, PCA chooses the direction along which the variance of the projected data is maximized.
The quality of the approximation worsens with the given data deviating from a linear relationship (when we want to fit a straight line through the data).
In the worst case, consider the data distributed around a circle.
In this case, all the directions through the mean will have approximately the same variance of the projected data, thus there is no "best" direction.
In the first case, there appears to be a linear relationship in the given data, thus reducing the dimension

to 1 using PCA yields good quality results.
In the second case, the data is distributed in an approximate ellipse, thus the major axis will be the direction which maximizes the variance of the projected data, however, in case of an ellipse, there is considerable variance along the minor axis too, PCA approximation loses this information about the original data and thus the quality of results is poor compared to the first set of data.
Thus, trying to reduce the dimension of the data to 1, in data which varies along more than one direction is meaningless in most of the cases.

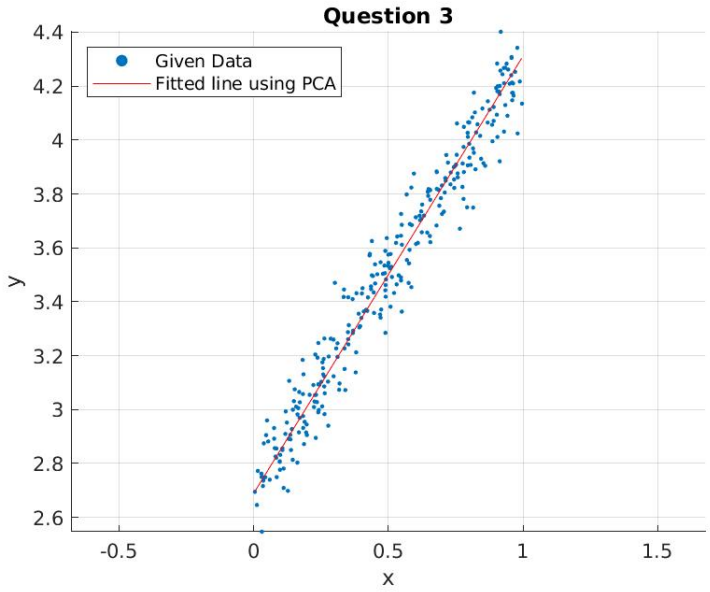

Figure 1: For the first dataset

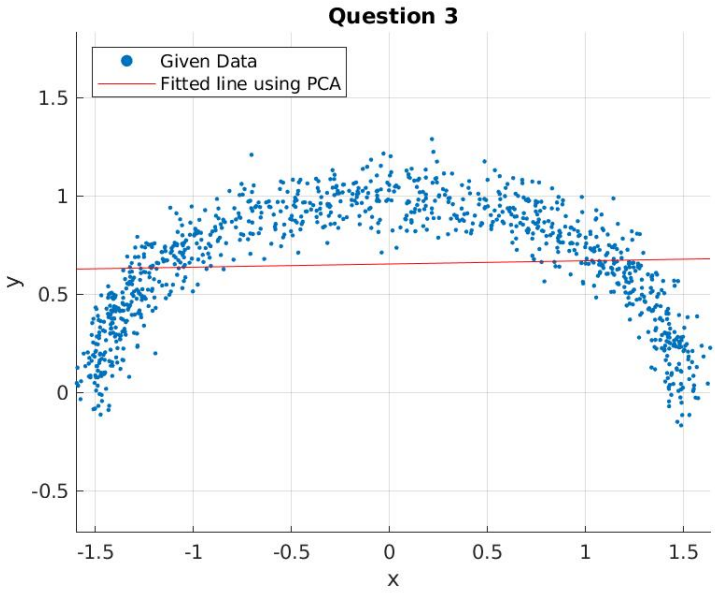

Figure 2: For the second dataset