ENPM673 - Homework1

Submitted by:

Srujan Panuganti	116302319
Mrinali Vyas	116189866
Akshitha Pothamshetty	116399326

Eigen Values/ Eigen Vectors:

1. Eigen Vectors of a transformation T, can be perceived as the vectors that doesn't change when the Transformation T is applied. Whereas, Eigen values are the magnitude of the Eigen vectors.
2. They play a role in dimensionality reduction. Occurence of dependent eigen vectors for a transformation matrix reduces the dimension after applying the transformation.
3. In our problem, we have obtained the eigen values and vectors for the covariance matrix. These eigen value decomposition helps us in representing the transformation matrix from the white data to the observational data.
4. The eigen values affect the scaling of data, whereas, the largest eigen vector show the direction of the data.
5. The eigen decomposition is used in applications involving Linear regression, Principle component analysis

Least square method:

Least square method helps us to find the best fit line for the set of data points and gives us the relationship between the data points. The independent variables are plotted on the X-axis and the dependent variables on Y-axis, this plotting will form the equation of the line by reducing the sum of offsets of points.

Steps:-

We have followed the below steps to achieve the task:

1. We have prepared the given data in the below format

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix} \quad B = \begin{bmatrix} m \\ b \end{bmatrix}$$

2. We have implemented the below formula using the prepared data

$$B = (X^T X)^{-1}(X^T Y)$$

3. The matrix B gives the slope and the y-intercept to form the line equation of best fit using the least squares method.

Problems/ Solutions:-

1. The problem of the least squares is that it also considers the observational error in the dependent variables and not in the independent variables.
2. Any outliers in the data could lead to bad line fitting.

Total Least square:

It is a modelling technique which accounts both the dependent and the independent errors in the observational data.

Steps:-

In our algorithm, we have followed the following steps to accomplish the task:

1. We have used the perpendicular distance based least square technique to accomplish this task.
2. We have prepared the data as given below:

$$\begin{bmatrix} x_1 - \bar{x} & y_1 - \bar{y} \\ \vdots & \vdots \\ x_n - \bar{x} & y_n - \bar{y} \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}$$

3. Using the above prepared data we have obtained the A_T_A matrix.
4. We found the eigen values and the eigen vectors of the A_T_A matrix.
5. The eigen vector corresponding to the smallest eigen value calculated from the A_T_A matrix gives the points for the line equation which is the solution of the total least squares technique.
6. We have obtained the slope from the best fit set of points and then plotted the line.

Observations:

1. As both the variables are measured in the same units the errors on them represent the shortest distance between the data point and the fitted curve.

OUTLIERS

They are data values that differ from the remaining data which means they are either erroneous or have been generated from other external source.

We have chosen to use the Random Sample Consensus(RANASC) Technique for Outlier Rejection. The main reason for choosing this technique over others is, the RANSAC algorithm works with high data variances as we can easily reject the outliers using a threshold.

RANSAC

It is an outlier detection method which estimates the parameters of a model from a set of given data points which contain outliers, these outliers are data which does not influence the estimate value. Its main purpose is to model the inliers from the given data.

Working-

As the given data is 2D, 2 points are sufficient to fit the model. The algorithm randomly selects 2 points and assumes them to be inliers and tests the other points to see if it fits the model. If the model can classify sufficient point as inliers then the algorithm is optimum.
This procedure is iterated a times for the most fitted model of inliers.

Steps:-

In our algorithm, we have followed the following steps to accomplish the task

1. Defined a function to calculate the perpendicular distance from a point to a line model.
2. Selected two random points using the the numpy ‘randint’ function
3. Calculated the slope and y-intercept to form a line model
4. Used the line model and the perpendicular distance function to compare with a threshold value.
5. Segregated the inliers and outliers based on the threshold value.
6. This process is run in the loop until a inlier ratio criterion is met.

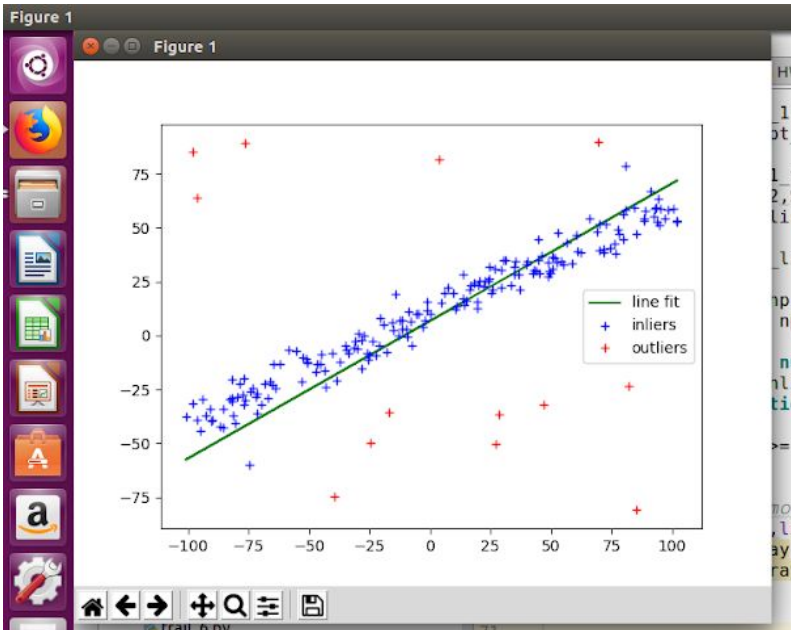
Observations:

1. Different threshold values and inlier ratio criterion are experimented for different sets of data that is provided.
2. When, the same threshold and inlier ratio is maintained for the three datasets, the data1 took least number of steps to achieve the inlier ratio criterion. The data3 took the highest number of iterations to achieve the criterion.
3. To regulate the number of iteration, different target inlier ratios are given to different datasets.
4. Increase in the number of iterations decreased the number of inliers detected

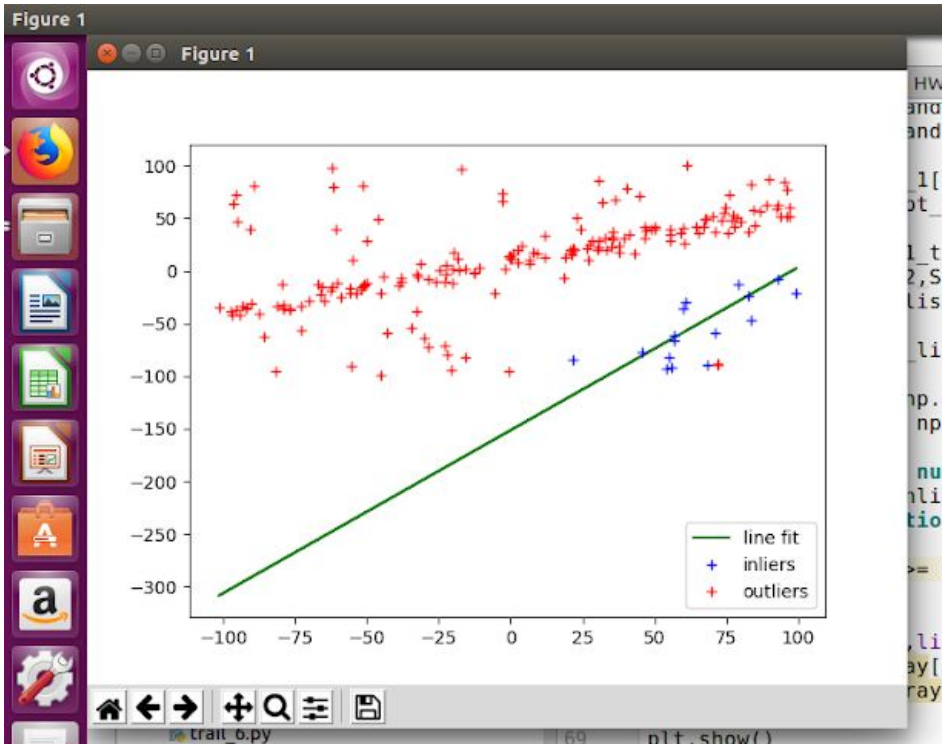
Results:

Below are the results plotted for different datasets for same inlier ratio = 0.9

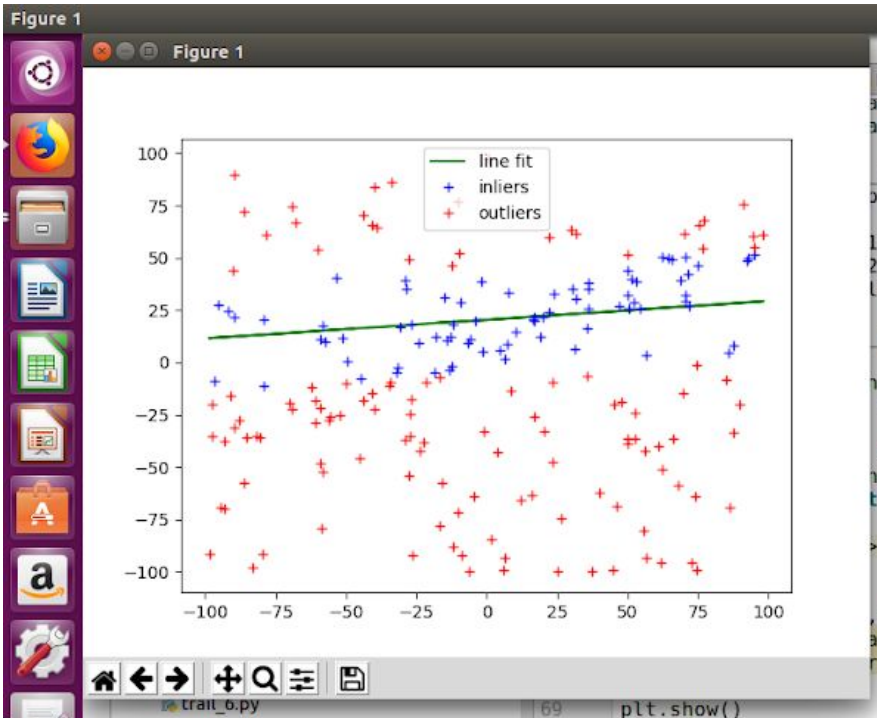
For Dataset 1:
Inlier ratio:0.9
Number of iterations: 2



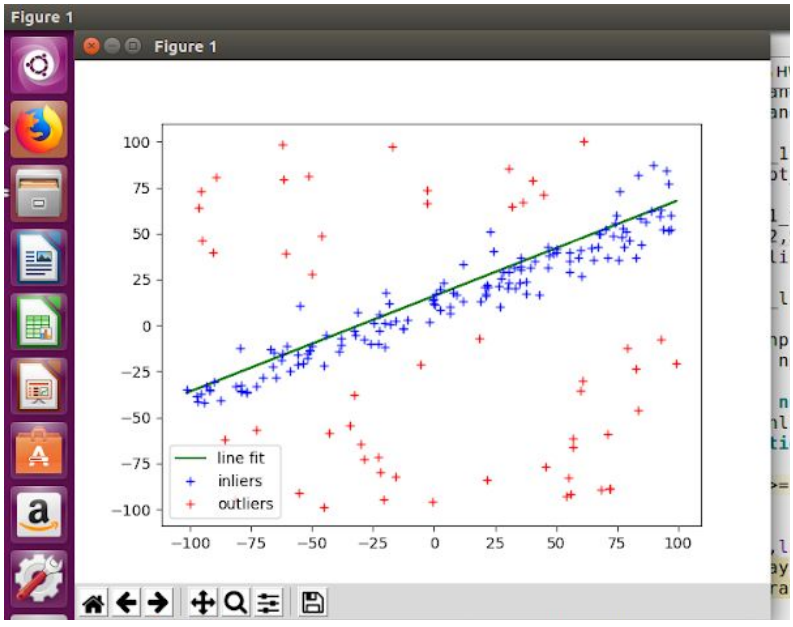
For dataset 2:
Inlier ratio:0.9
Number of iterations: maximum



For dataset 3:
Inlier ratio:0.9
Number of iterations: maximum



Below are the plot for optimized RANSAC algorithm
For dataset2:
Inlier ratio:0.7
Number of iterations: 16



For dataset3:
Inlier ratio:0.5
Number of iterations:9