Carbon Nanotubes Data Set Analysis and Experiments

Project Report (Statistical Data Analysis)

Rajat Singh {rajat.s15@iiits.in}
Tanmay Kalani {tanmay.k16@iiits.in}

1. Abstract

In this report, three regression models [i.e Linear Regression (LR), Support Vector Regression (SVR) and Multi-Layer Perceptron Regression (MLPR)] have been used for atomic coordinate prediction of carbon nanotubes (CNTs). The research reported in this study has two primary objectives: (1) to test these three prediction models that calculate atomic coordinates of CNTs instead of using any simulation software and (2) to compare the performance of these models using methods of Statistical Data Analysis.

2. Approach

2.1 Introduction

Carbon nanotubes (CNTs) have been introduced as the alternatives for copper/aluminum metallic interconnects to overcome problems caused from miniaturization. CNTs are 2-D graphene crystal as rolled-up sheets. Initial coordinates of all carbon atoms are generated randomly. Different chiral vectors are used for each CNT simulation. The atom type is selected as carbon, bond length is used as 1.42 AÂ° (default value). CNT calculation parameters are used as default parameters. Dataset Description is as follows:

- **Dataset Characteristics:** Univariate
- **Associated Tasks:** Regression
- **Missing Values:** N/A
- **Number of Instances:** 10721
- **Number of Attributes:** 8

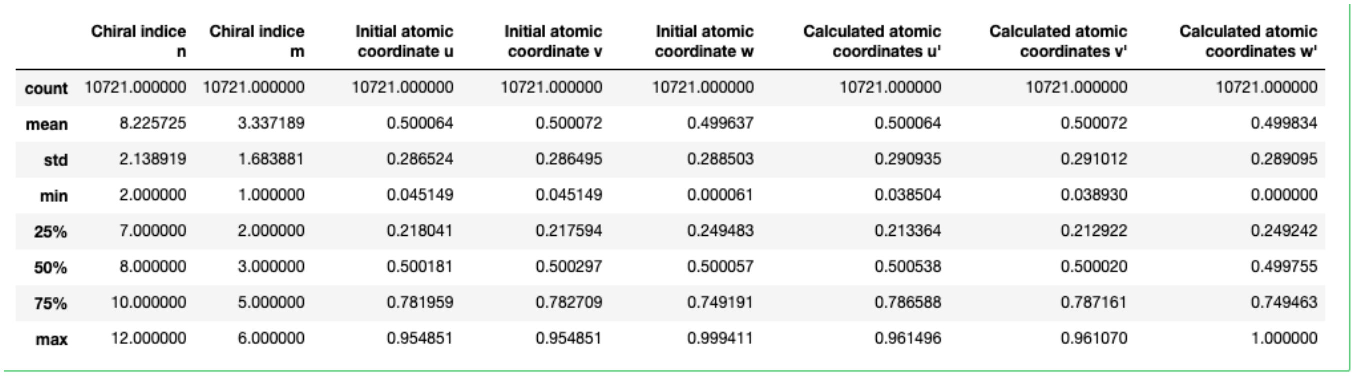
There are total eight attributes named as 1. Chiral indice n 2. Chiral indice m 3. Initial atomic coordinate u 4. Initial atomic coordinate v 5. Initial atomic coordinate w 6. Calculated atomic coordinate uâ€ 7. Calculated atomic coordinate vâ€ 8. Calculated atomic coordinate wâ€.

2.2 Methodology

The whole methodology involves seven subsections:

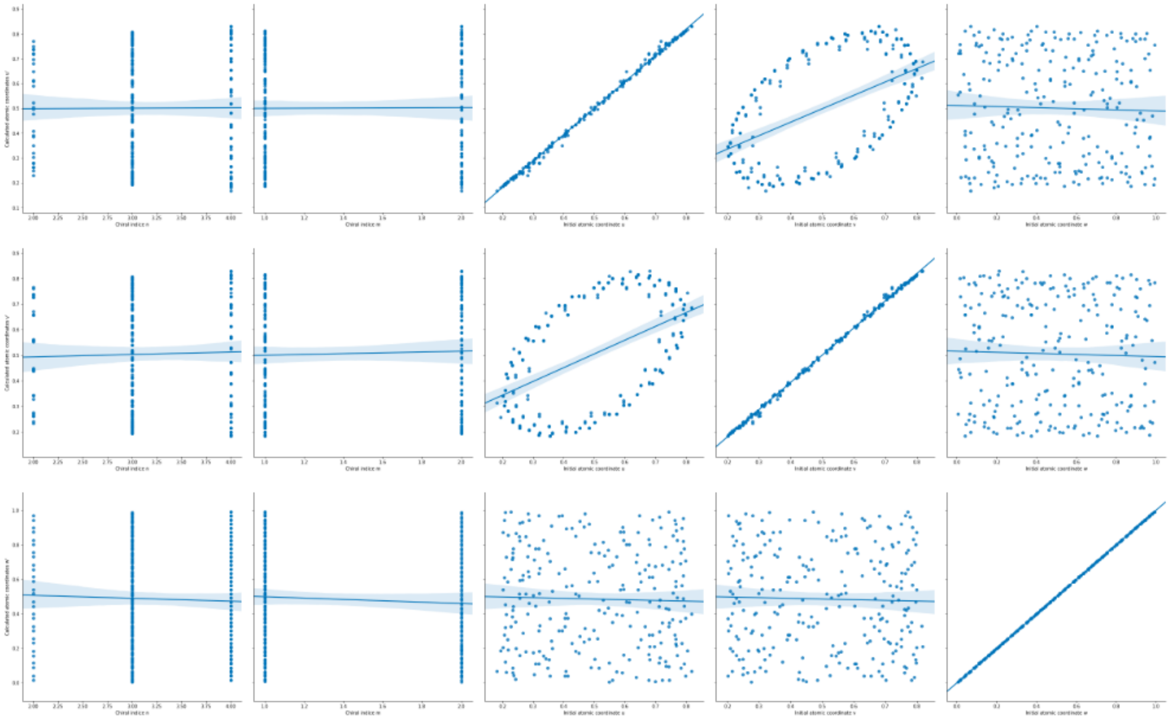
- Statistical Analysis
- Non-collinearity Check
- Principal Component Analysis
- Linear Regression Model Test (with and without PCA)
- Model Adequacy Test
- Support Vector Regression
- Multilayer Perceptron Regression

1. Statistical Analysis



	Chiral indice n	Chiral indice m	Initial atomic coordinate u	Initial atomic coordinate v	Initial atomic coordinate w	Calculated atomic coordinates u'	Calculated atomic coordinates v'	Calculated atomic coordinates w'
count	10721.000000	10721.000000	10721.000000	10721.000000	10721.000000	10721.000000	10721.000000	10721.000000
mean	8.225725	3.337189	0.500064	0.500072	0.499637	0.500064	0.500072	0.499834
std	2.138919	1.683881	0.286524	0.286495	0.288503	0.290935	0.291012	0.289095
min	2.000000	1.000000	0.045149	0.045149	0.000061	0.038504	0.038930	0.000000
25%	7.000000	2.000000	0.218041	0.217594	0.249483	0.213364	0.212922	0.249242
50%	8.000000	3.000000	0.500181	0.500297	0.500057	0.500538	0.500020	0.499755
75%	10.000000	5.000000	0.781959	0.782709	0.749191	0.786588	0.787161	0.749463
max	12.000000	6.000000	0.954851	0.954851	0.999411	0.961496	0.961070	1.000000

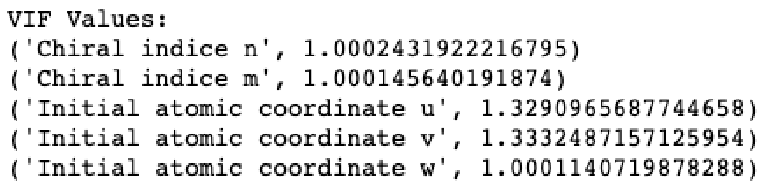
```
Out[22]: <seaborn.axisgrid.PairGrid at 0x12291d828>
```



In above fig,

First row describes the data distribution between u’ and (n, m, u, v and v respectively)
Second row describes the data distribution between v’ and (n, m, u, v and v respectively)
Third row describes the data distribution between w’ and (n, m, u, v and v respectively)

2. Non-Collinearity Check using **Variational Inflation Factor**

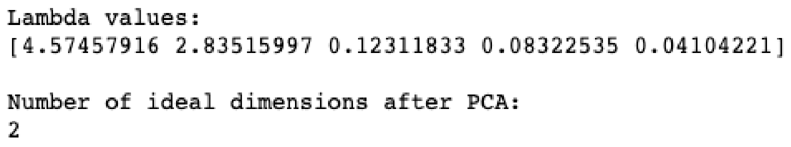


```
VIF Values:
('Chiral indice n', 1.0002431922216795)
('Chiral indice m', 1.000145640191874)
('Initial atomic coordinate u', 1.3290965687744658)
('Initial atomic coordinate v', 1.3332487157125954)
('Initial atomic coordinate w', 1.0001140719878288)
```

Since all the VIF values are less than 10, therefore all the features are desirable.

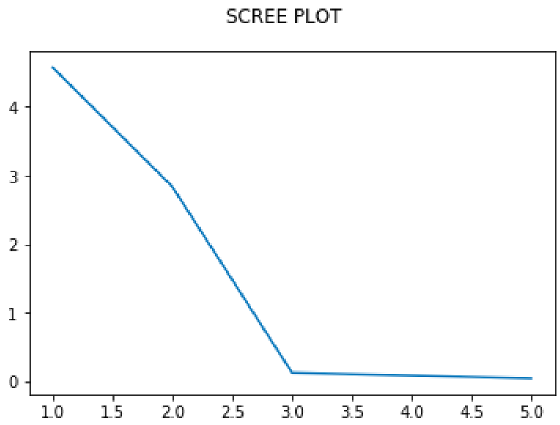
3. Principal Component Analysis

1. Using Average Root Method (i.e. retaining all Lambda(j) such that **Lambda(j) > Mean Lambda**)



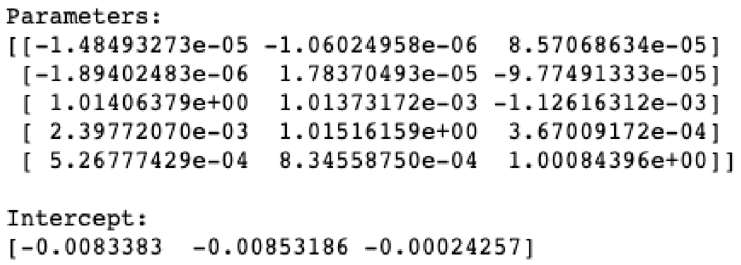
```
Lambda values:
[4.57457916 2.83515997 0.12311833 0.08322535 0.04104221]

Number of ideal dimensions after PCA:
2
```

2. Scree Plot



Here the elbow point is 3, therefore should retain all lambda(j) where (j)<=3

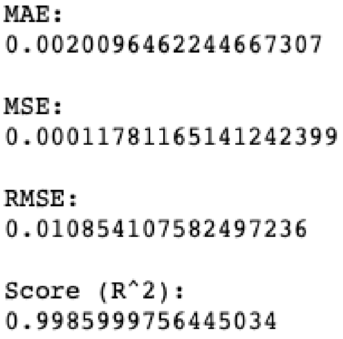
4 and 5. Linear Regression Model with Model Adequacy Test



```
Parameters:
[[-1.48493273e-05 -1.06024958e-06  8.57068634e-05]
 [-1.89402483e-06  1.78370493e-05 -9.77491333e-05]
 [ 1.01406379e+00  1.01373172e-03 -1.12616312e-03]
 [ 2.39772070e-03  1.01516159e+00  3.67009172e-04]
 [ 5.26777429e-04  8.34558750e-04  1.00084396e+00]]

Intercept:
[-0.0083383  -0.00853186 -0.00024257]
```

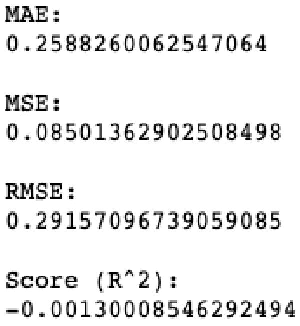
Above fig. shows the parameters value calculated using Linear Regression Model.

a. Without PCA



```
MAE:
0.0020096462244667307

MSE:
0.00011781165141242399

RMSE:
0.010854107582497236

Score (R^2):
0.9985999756445034
```

b. With PCA

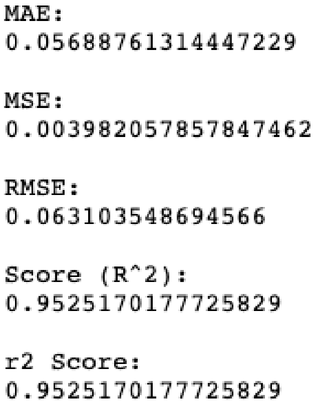


```
MAE:
0.2588260062547064

MSE:
0.08501362902508498

RMSE:
0.29157096739059085

Score (R^2):
-0.00130008546292494
```

Since R^2 value for (without PCA) is closer to 1, therefore model can be considered a good model. But in comparison, (with PCA) is not a good model.

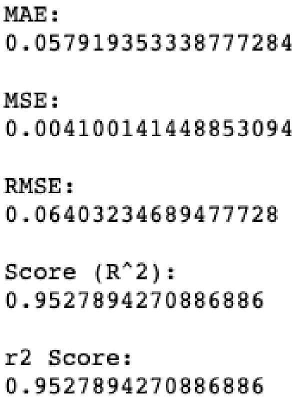
6. Support Vector Regression

a. Response var(u')



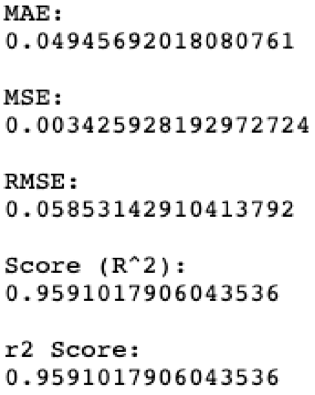
```
MAE:
0.05688761314447229

MSE:
0.003982057857847462

RMSE:
0.063103548694566

Score (R^2):
0.9525170177725829

r2 Score:
0.9525170177725829
```

b. Response var(v')



```
MAE:
0.057919353338777284

MSE:
0.004100141448853094

RMSE:
0.06403234689477728

Score (R^2):
0.9527894270886886

r2 Score:
0.9527894270886886
```

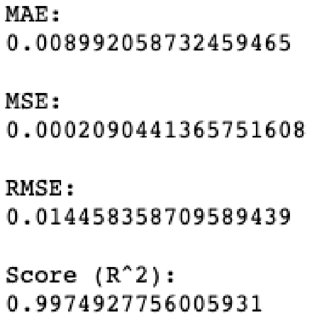
c. Response var(w')



```
MAE:
0.04945692018080761

MSE:
0.003425928192972724

RMSE:
0.05853142910413792

Score (R^2):
0.9591017906043536

r2 Score:
0.9591017906043536
```

In above fig. accuracy measures and R^2 scores are calculated for each response variables u', v', and w' respectively.

7. Multilayer Perceptron Regression



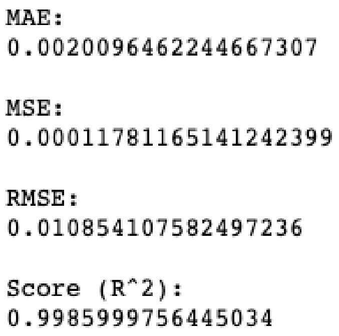
```
MAE:
0.008992058732459465

MSE:
0.0002090441365751608

RMSE:
0.014458358709589439

Score (R^2):
0.9974927756005931
```

3. Results and Conclusion

From above experiments, we found out that order of performance of these three models is as follow:

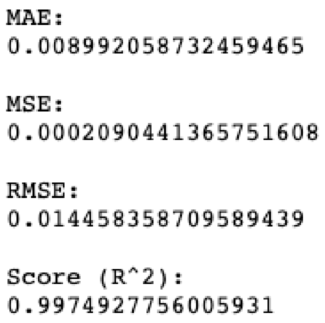
LR > MLPR > SVR

a. LR



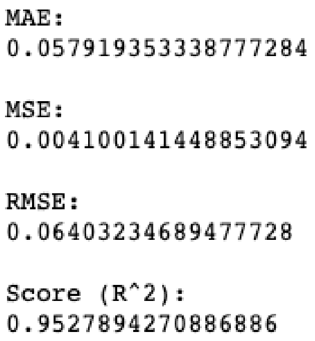
```
MAE:
0.0020096462244667307

MSE:
0.00011781165141242399

RMSE:
0.010854107582497236

Score (R^2):
0.9985999756445034
```

b. MLPR



```
MAE:
0.008992058732459465

MSE:
0.0002090441365751608

RMSE:
0.014458358709589439

Score (R^2):
0.9974927756005931
```

c. SVR



```
MAE:
0.057919353338777284

MSE:
0.004100141448853094

RMSE:
0.06403234689477728

Score (R^2):
0.9527894270886886
```

Through our analysis, we concluded that, this problem of prediction of atomic coordinates of CNTs is a Regression problem which can be easily solved using Linear Regression Model**.**

4. Related work and References

1. ACI, M, AVCI, M. (2016). ARTIFICIAL NEURAL NETWORK APPROACH FOR ATOMIC COORDINATE PREDICTION OF CARBON NANOTUBES. Applied Physics A, 122, 631
2. https://scikit-learn.org/
3. https://www.geeksforgeeks.org
4. https://archive.ics.uci.edu/ml/datasets/Carbon+Nanotubes