Dataset2_Friedman1_output_8

October 20, 2021

1 Dataset 2 - Friedman 1

1.1 Experiment Details

The aim of the experiment is to verify if the: 1. ABC_GAN model corrects model misspecification 2. ABC_GAN model performs better and converges faster than a simple C-GAN model

In the experiment we predict the distribution that represents the real data and simulate realistic fake data points using statistical mode, C-GAN and ABC-GAN model with 3 priors. We analyze and compare their performance using metrics like mean squared error, mean absolute error , manhattan distance and euclidean distance between y_{real} and y_{pred}

The models are as follows:

1. The statistical model assumes the distribution $Y = \beta X + \mu$ where $\mu \sim N(0,1)$
2. The Conditional GAN consists of
   1. Generator with 2 hidden layers with 100 nodes each and ReLu activation.
   2. Discriminator with 2 hidden layers with 25 and 50 nodes and ReLu activation. We use Adam's optimser and BCE Logit Loss to train the model. The input to the Generator of the GAN is (x,e) where x are the features and $e \sim N(0,1)$. The discriminator output is linear.
3. The ABC GAN Model consists of
   1. ABC generator is defined as follows:
      1. $Y = 1 + \beta_1 x_1 + \beta_2 x_2 + \beta_2 x_3 + ... + \beta_n x_n + N(0,\sigma)$ where $\sigma = 0.1$
      2. $\beta_i \sim N(0,\sigma^*)$ when $\mu = 0$ else $\beta_i \sim N(\beta_i^*, \sigma^*)$ where $\beta_i^* s$ are coefficients obtained from statistical model
      3. σ^* takes the values 0.01,0.1 and 1

   2. C-GAN network is as defined above. However the input to the Generator of the GAN is (x, y_{abc}) where y_{abc} is the output of the ABC Generator.

1.2 Import Libraries

[1]:
```
import warnings
warnings.filterwarnings('ignore')
```

[2]:
```
import train_test
import ABC_train_test
import regressionDataset
import network
```

```
import statsModel
import performanceMetrics
import friedman1Dataset
import dataset
import sanityChecks
import torch
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import norm
from torch.utils.data import Dataset,DataLoader
from torch import nn
```

1.3 Parameters

General Parameters

1. Number of Samples
2. Number of features

ABC-Generator parameters are as mentioned below: 1. mean : 1 ($\beta \sim N(\beta^*, \sigma)$ where β^* are coefficients of statistical model) or 1 ($\beta \sim N(0, \sigma)$ 2. std : $\sigma = 1, 0.1, 0.01$ (standard deviation)

[3]:
```
n_features = 10
n_samples= 100

#ABC Generator Parameters
mean = 1
variance = 0.001
```

[4]:
```
# Parameters
n_samples = 100
n_features = 10
mean = 1
variance = 0.01
```

1.4 Dataset

Friedman 1 Dataset

- $y(X) = 10 * sin(pi * X_0 * X_1) + 20 * (X_2 - 0.5) * *2 + 10 * X_3 + 5 * X_4 + noise * N(0, 1)$.
- Only 5 features used to calculate y
- Noise is Gaussian
- 1000 datapoints and 10 features used in the following experiment

[5]:
```
X, Y = friedman1Dataset.friedman1_data(n_samples,n_features)
```

```
         X0        X1        X2        X3        X4        X5        X6  \
0  0.205399  0.007976  0.796151  0.775126  0.731206  0.768476  0.327187
1  0.760839  0.645001  0.593680  0.224939  0.154804  0.886347  0.733311
2  0.528509  0.103180  0.833456  0.999879  0.886985  0.527743  0.497971
```

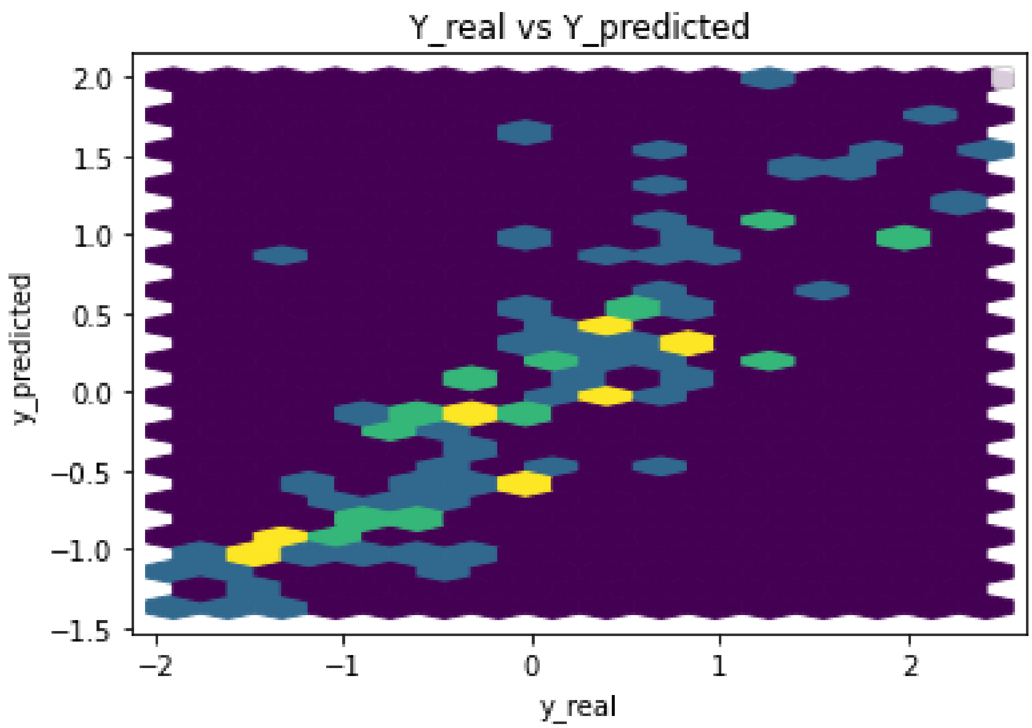
```
3  0.554852  0.825196  0.462980  0.041245  0.304671  0.438636  0.106507
4  0.994753  0.950638  0.344231  0.204365  0.912542  0.118774  0.358425

         X7        X8        X9          Y
0  0.887998  0.776688  0.036306  13.307025
1  0.924330  0.043927  0.546543  13.279826
2  0.705584  0.702099  0.363390  18.349689
3  0.137886  0.689228  0.164774  12.085012
4  0.219459  0.150113  0.174542   8.848023
```

1.5 Stats Model

[6]:
```
[coeff,y_pred] = statsModel.statsModel(X,Y)
```

```
No handles with labels found to put in legend.

                            OLS Regression Results
==============================================================================
Dep. Variable:                      Y   R-squared:                       0.689
Model:                            OLS   Adj. R-squared:                  0.654
Method:                 Least Squares   F-statistic:                     19.67
Date:                Wed, 20 Oct 2021   Prob (F-statistic):           1.36e-18
Time:                        20:18:26   Log-Likelihood:                -83.570
No. Observations:                 100   AIC:                             189.1
Df Residuals:                      89   BIC:                             217.8
Df Model:                          10
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       2.526e-15      0.059   4.27e-14      1.000      -0.118       0.118
x1             0.4379      0.065      6.726      0.000       0.309       0.567
x2             0.3931      0.067      5.854      0.000       0.260       0.526
x3             0.1223      0.064      1.921      0.058      -0.004       0.249
x4             0.5993      0.061      9.848      0.000       0.478       0.720
x5             0.2367      0.061      3.886      0.000       0.116       0.358
x6             0.0365      0.061      0.603      0.548      -0.084       0.157
x7             0.0412      0.065      0.632      0.529      -0.088       0.171
x8            -0.0205      0.061     -0.335      0.738      -0.142       0.101
x9            -0.0280      0.065     -0.434      0.665      -0.156       0.100
x10            0.0771      0.061      1.268      0.208      -0.044       0.198
==============================================================================
Omnibus:                       13.962   Durbin-Watson:                   1.932
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               22.650
Skew:                          -0.593   Prob(JB):                     1.21e-05
Kurtosis:                       5.008   Cond. No.                         1.88
==============================================================================

Notes:
```

```
[1] Standard Errors assume that the covariance matrix of the errors is correctly
specified.
Parameters:  const    2.525757e-15
x1       4.378637e-01
x2       3.930608e-01
x3       1.222633e-01
x4       5.992604e-01
x5       2.366984e-01
x6       3.654514e-02
x7       4.119880e-02
x8      -2.046353e-02
x9      -2.801962e-02
x10      7.708504e-02
dtype: float64
```




```
Performance Metrics
Mean Squared Error: 0.311462334526192
Mean Absolute Error: 0.4141847166346712
Manhattan distance: 41.41847166346712
Euclidean distance: 5.580881064188629
```

1.6 Common Training Parameters (GAN & ABC_GAN)

[7]:
```
n_epochs = 5000
error = 0.001
batch_size = n_samples//2
```

1.7 GAN Model

[8]:
```
real_dataset = dataset.CustomDataset(X,Y)
device = torch.device('cuda' if torch.cuda.is_available() else 'cpu')
```

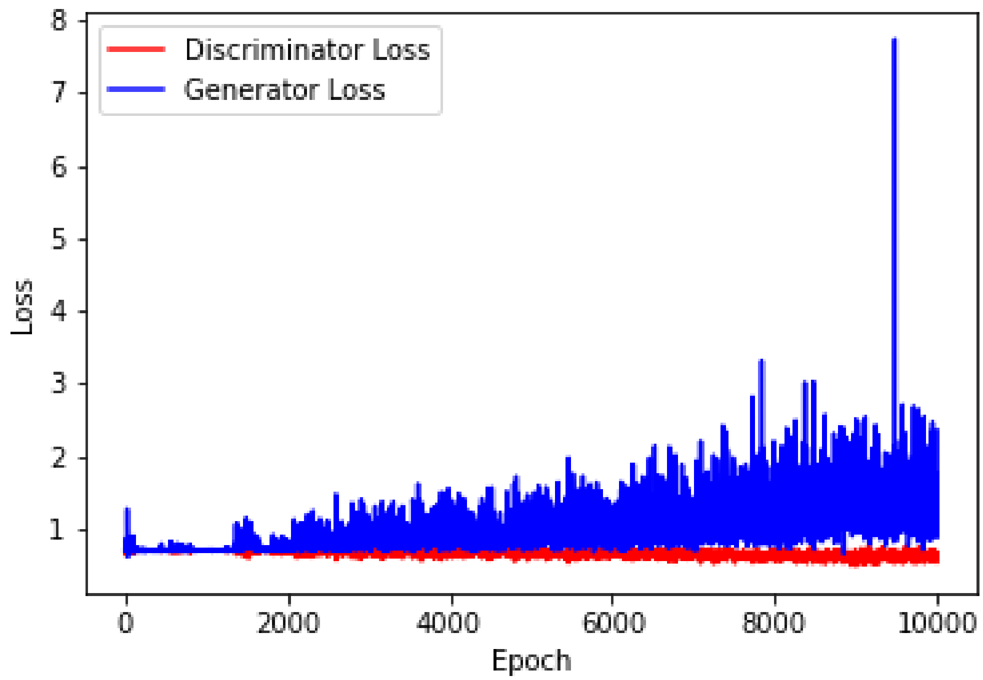
Training GAN for n_epochs number of epochs

[9]:
```
generator = network.Generator(n_features+2)
discriminator = network.Discriminator(n_features+2)

criterion = torch.nn.BCEWithLogitsLoss()
gen_opt = torch.optim.Adam(generator.parameters(), lr=0.01, betas=(0.5, 0.999))
disc_opt = torch.optim.Adam(discriminator.parameters(), lr=0.01, betas=(0.5, 0.
 ↪999))
```

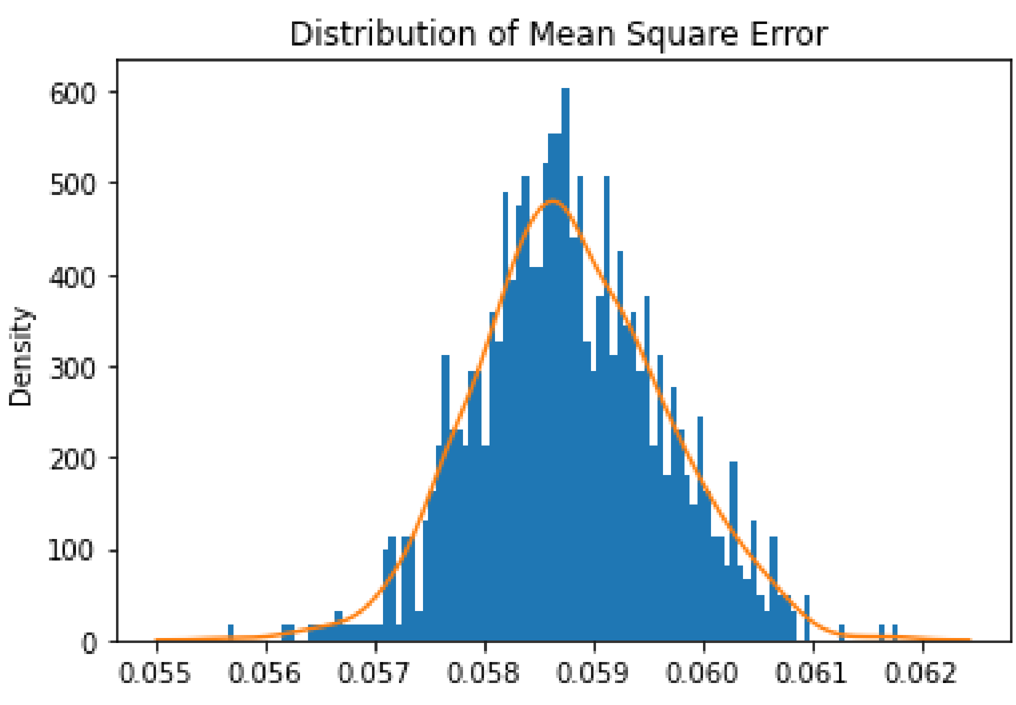
[10]:
```
print(generator)
print(discriminator)
```

```
Generator(
  (hidden1): Linear(in_features=12, out_features=100, bias=True)
  (hidden2): Linear(in_features=100, out_features=100, bias=True)
  (output): Linear(in_features=100, out_features=1, bias=True)
  (relu): ReLU()
)
Discriminator(
  (hidden1): Linear(in_features=12, out_features=25, bias=True)
  (hidden2): Linear(in_features=25, out_features=50, bias=True)
  (output): Linear(in_features=50, out_features=1, bias=True)
  (relu): ReLU()
)
```

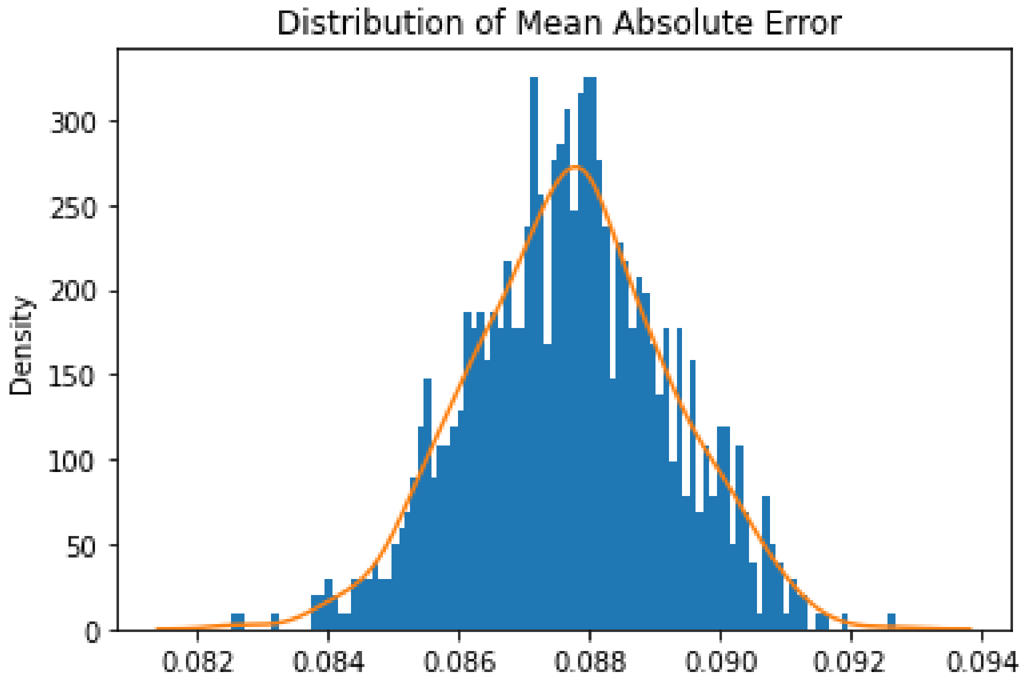
[11]:
```
train_test.
 ↪training_GAN(discriminator,generator,disc_opt,gen_opt,real_dataset,batch_size,␣
 ↪n_epochs,criterion,device)
```



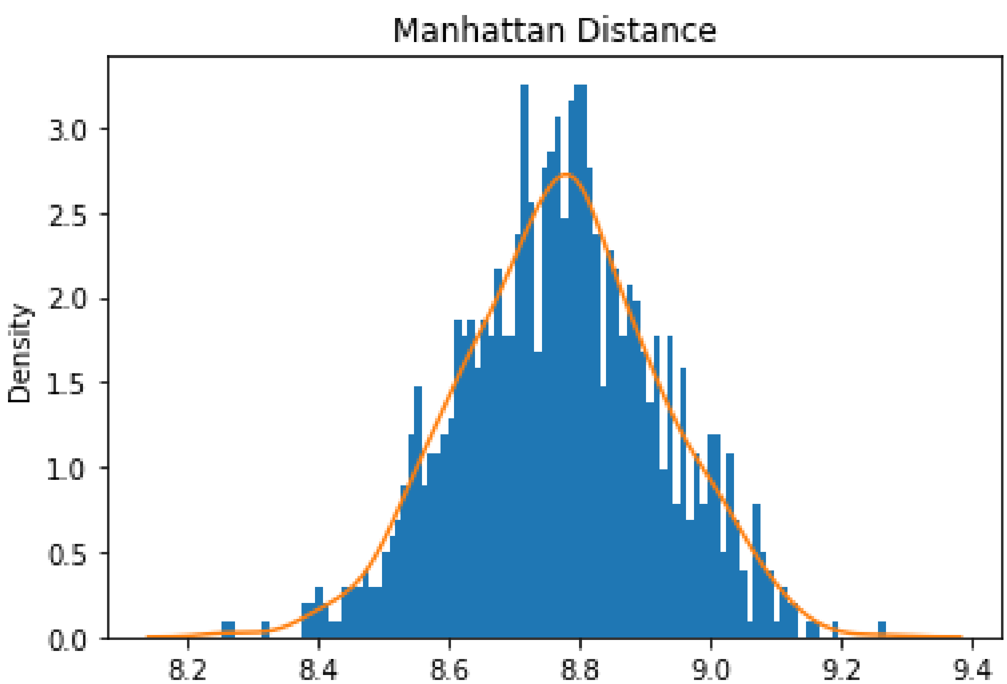
[12]:
```
train_test.test_generator(generator,real_dataset,device)
```



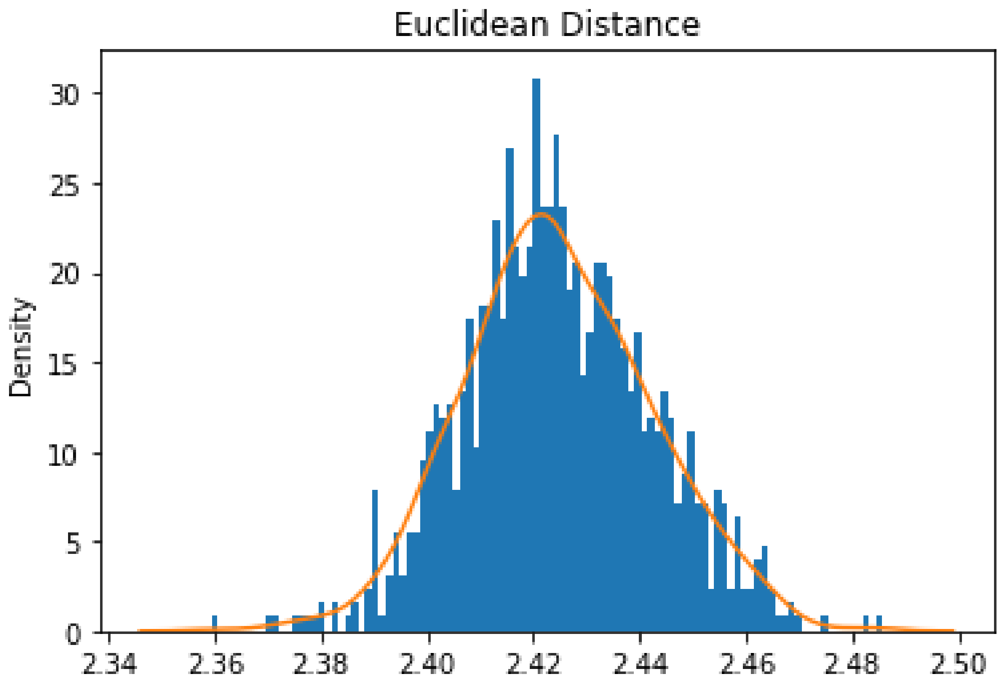
```
Mean Square Error: 0.05879006125789454
```




```
Mean Absolute Error: 0.0877062600217387
```

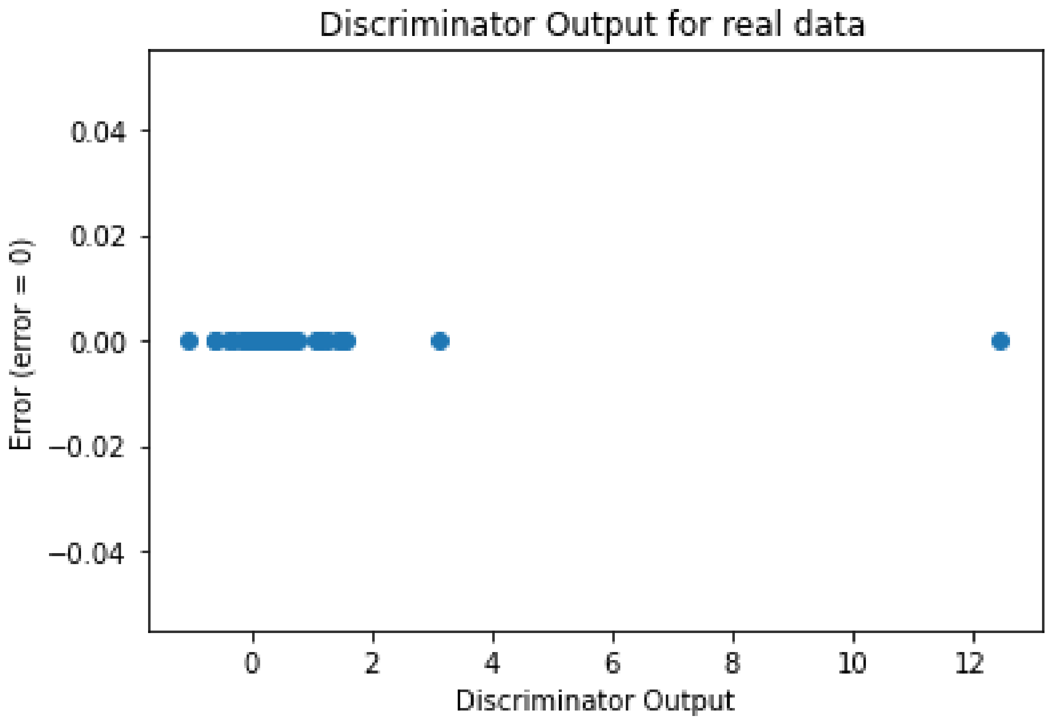


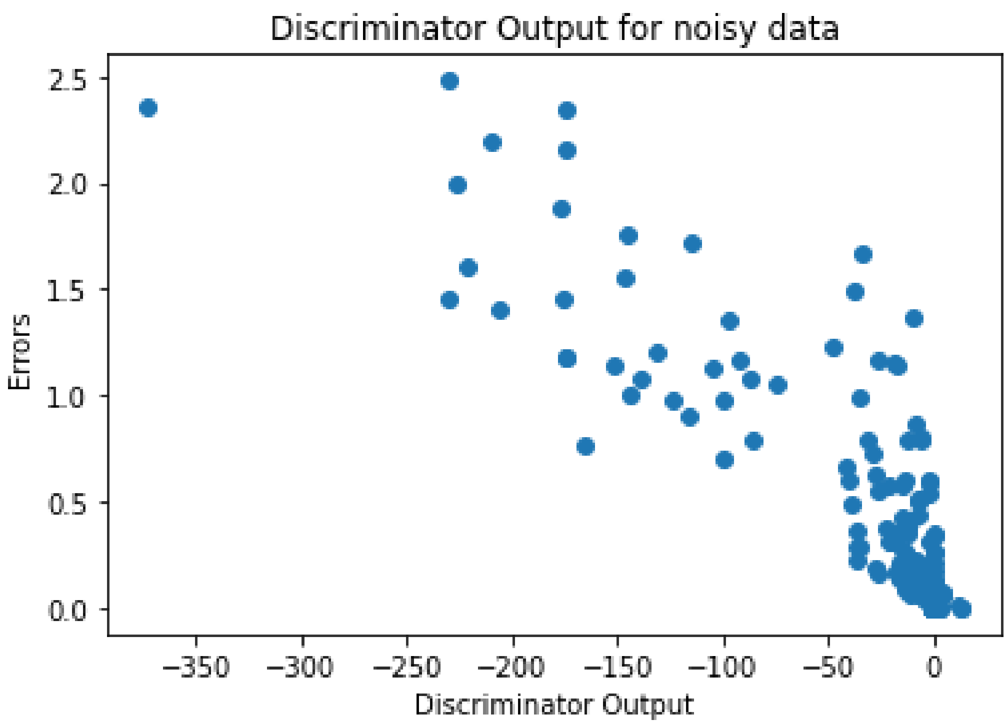
```
Mean Manhattan Distance: 8.77062600217387
```




```
Mean Euclidean Distance: 2.4246019282875513
```

[13]:
```
sanityChecks.discProbVsError(real_dataset,discriminator,device)
```


Discriminator Output for real data
0.04
0.02
0.00
−0.02
−0.04
Error (error = 0)
0
2
4
6
8
10
12
Discriminator Output

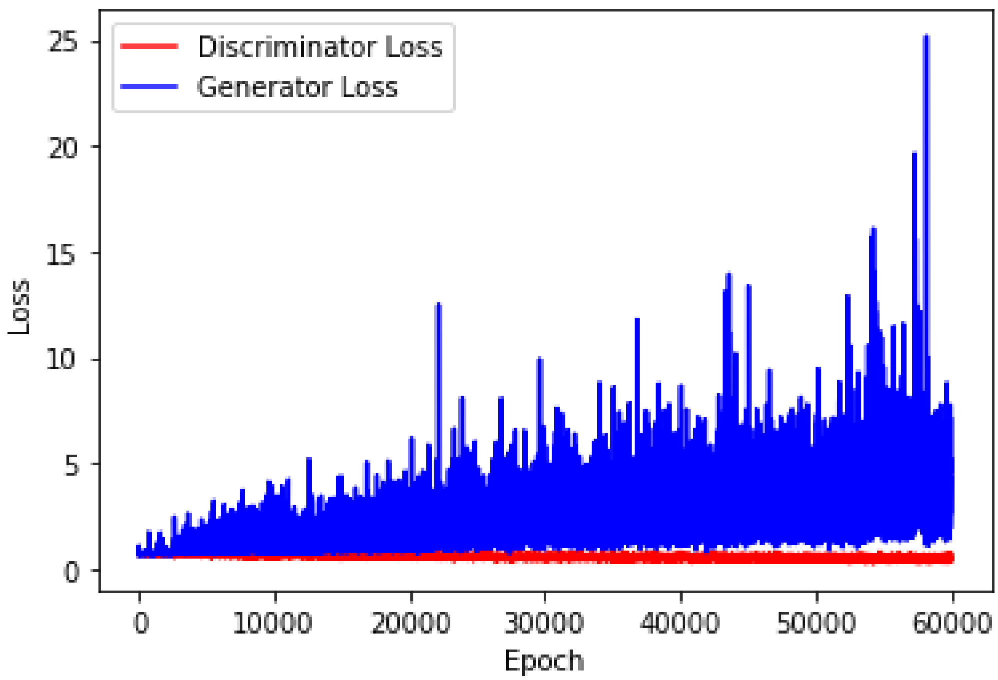

Discriminator Output for noisy data
2.5
2.0
1.5
1.0
0.5
0.0
Errors
−350
−300
−250
−200
−150
−100
−50
0
Discriminator Output

Training GAN until mse of y_pred is > 0.1 or n_epochs < 30000

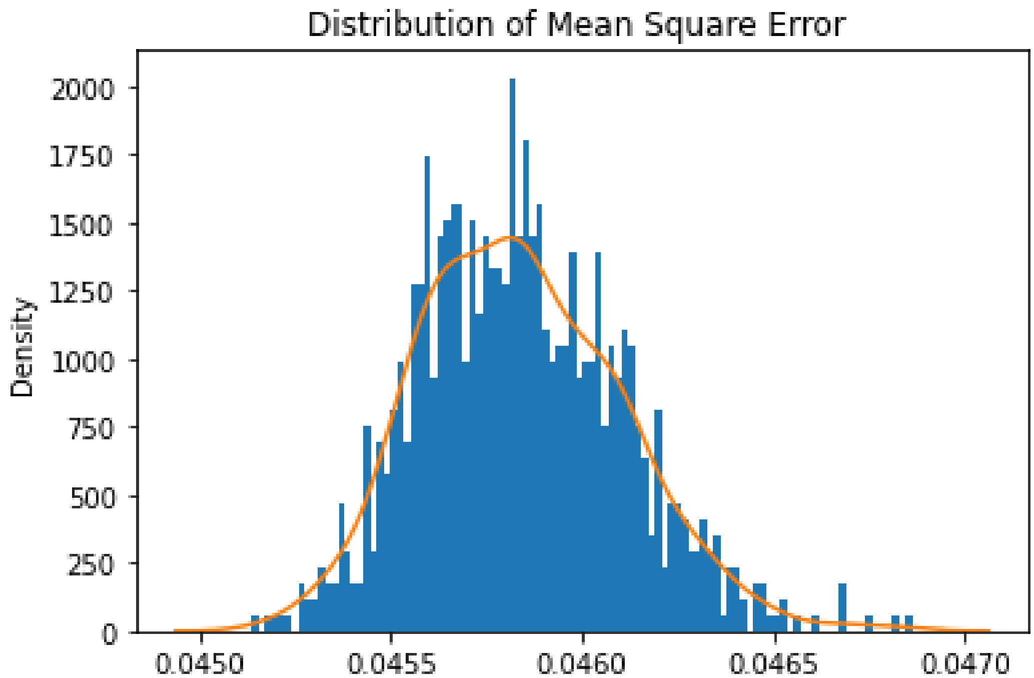
[14]:
```
generator = network.Generator(n_features+2)
discriminator = network.Discriminator(n_features+2)
criterion = torch.nn.BCEWithLogitsLoss()
gen_opt = torch.optim.Adam(generator.parameters(), lr=0.01, betas=(0.5, 0.999))
disc_opt = torch.optim.Adam(discriminator.parameters(), lr=0.01, betas=(0.5, 0.
 ↪999))
```

[15]:
```
train_test.
 ↪training_GAN_2(discriminator,generator,disc_opt,gen_opt,real_dataset,batch_size,error,crite
```

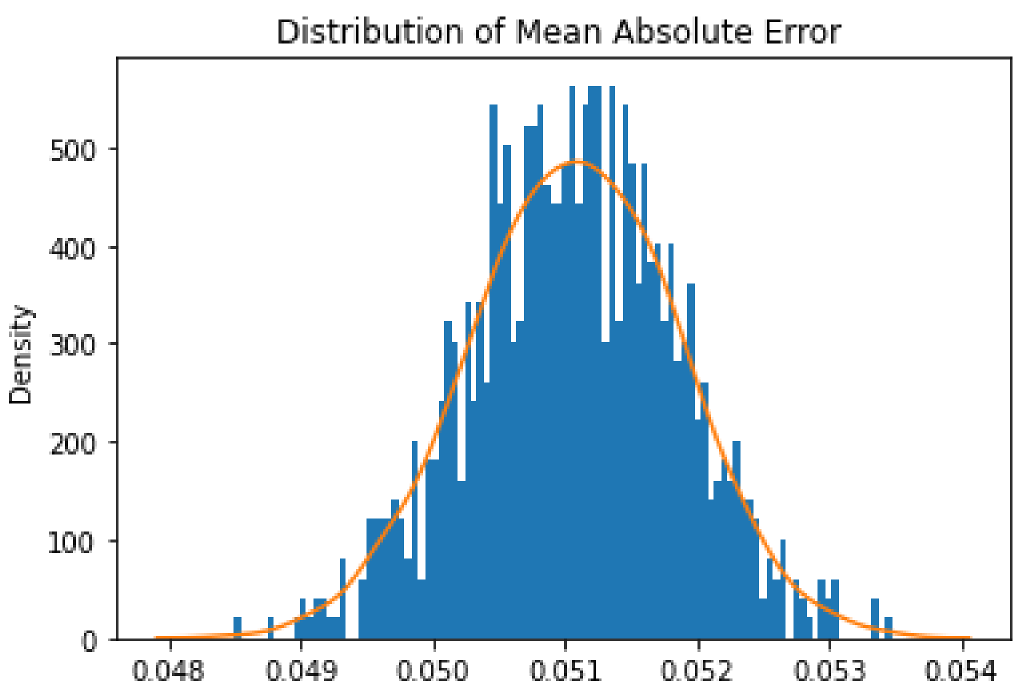
Number of epochs needed 30000



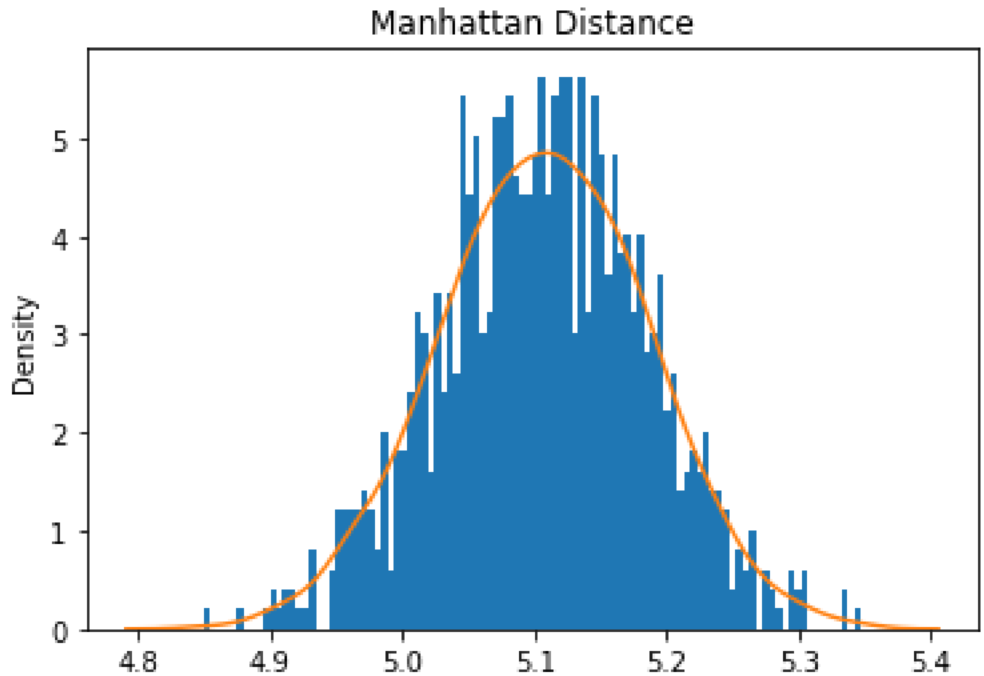
[16]:
```
train_test.test_generator(generator,real_dataset,device)
```



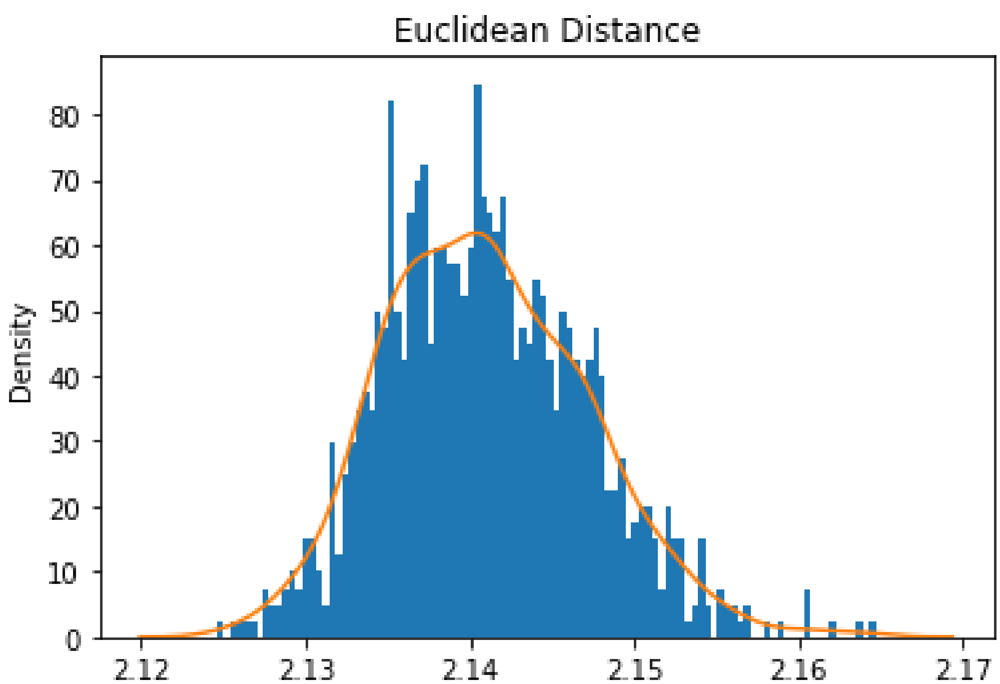

```
Mean Square Error: 0.045839322717389164
```



```
Mean Absolute Error: 0.05106582079738378
```




```
Mean Manhattan Distance: 5.106582079738379
```



Mean Euclidean Distance: 2.1410030736424503

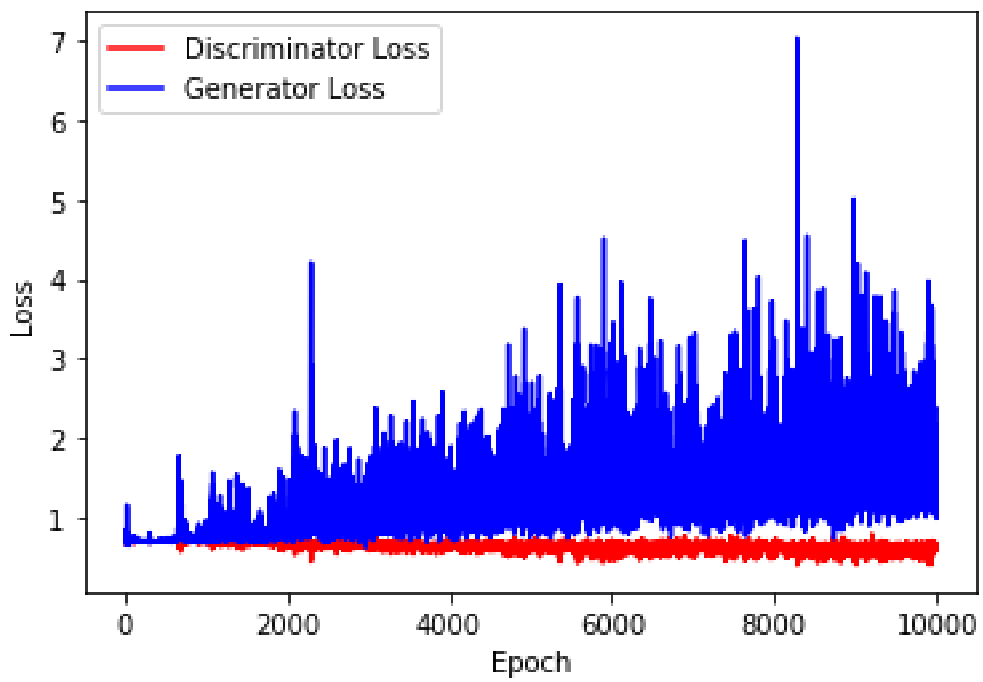
2 ABC GAN Model

2.0.1 Training the network

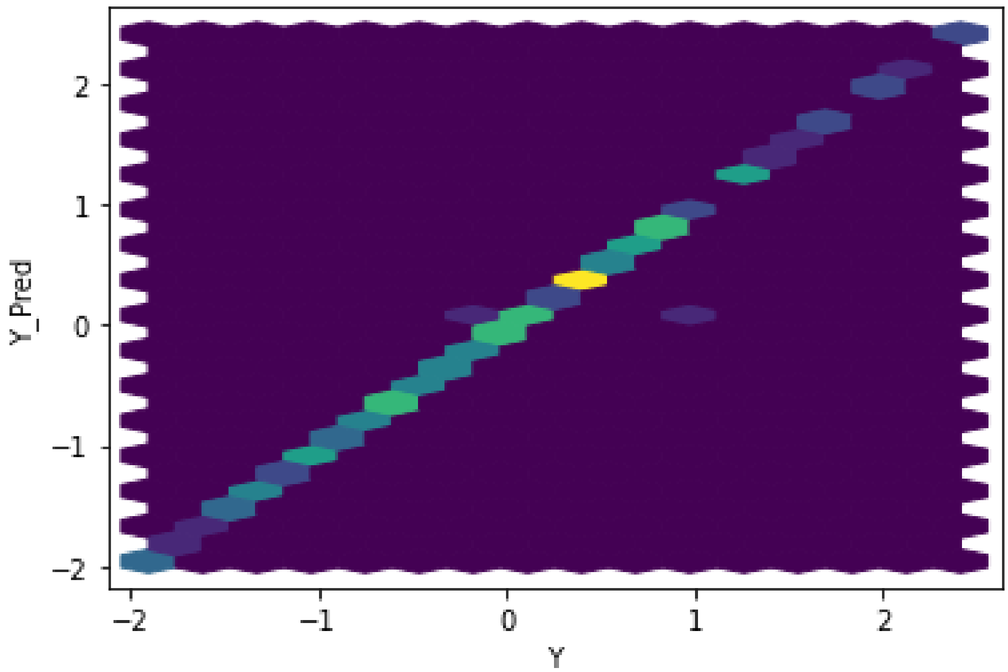
Training ABC-GAN for n_epochs number of epochs

[17]:
```
gen = network.Generator(n_features+2)
disc = network.Discriminator(n_features+2)

criterion = torch.nn.BCEWithLogitsLoss()
gen_opt = torch.optim.Adam(gen.parameters(), lr=0.01, betas=(0.5, 0.999))
disc_opt = torch.optim.Adam(disc.parameters(), lr=0.01, betas=(0.5, 0.999))
```

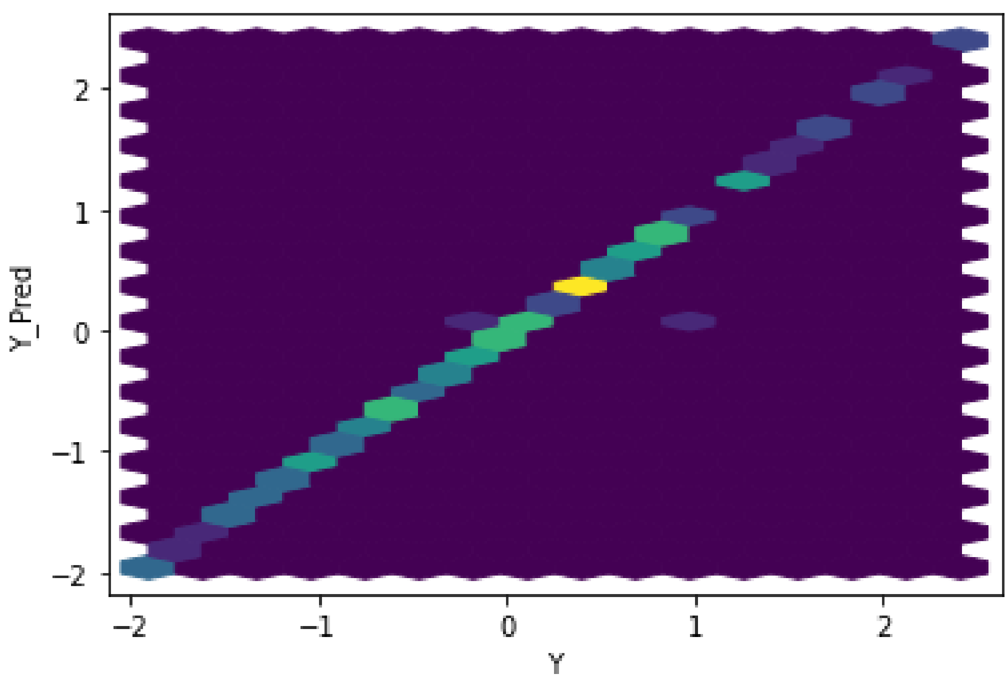
[18]:
```
ABC_train_test.training_GAN(disc, gen,disc_opt,gen_opt,real_dataset,␣
↪batch_size, n_epochs,criterion,coeff,mean,variance,device)
```



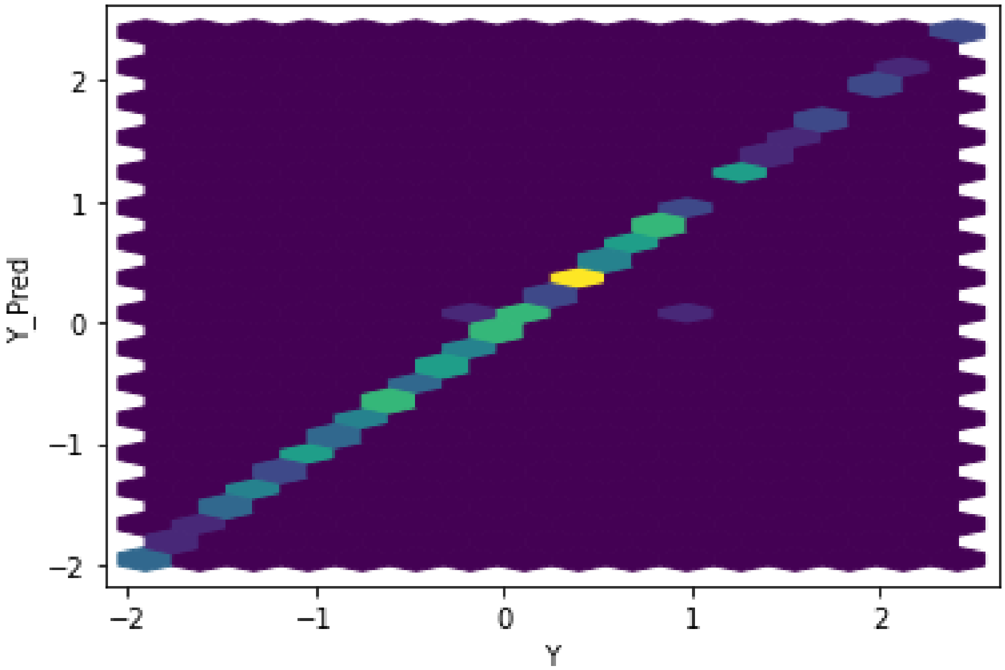
[19]:
```
ABC_train_test.test_generator(gen,real_dataset,coeff,mean,variance,device)
```

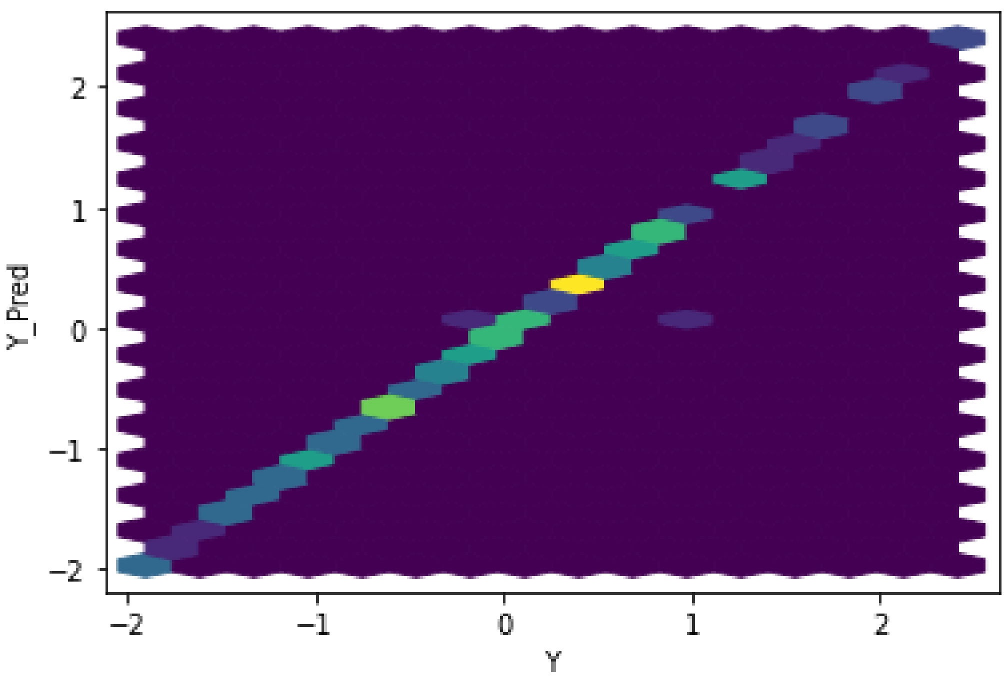


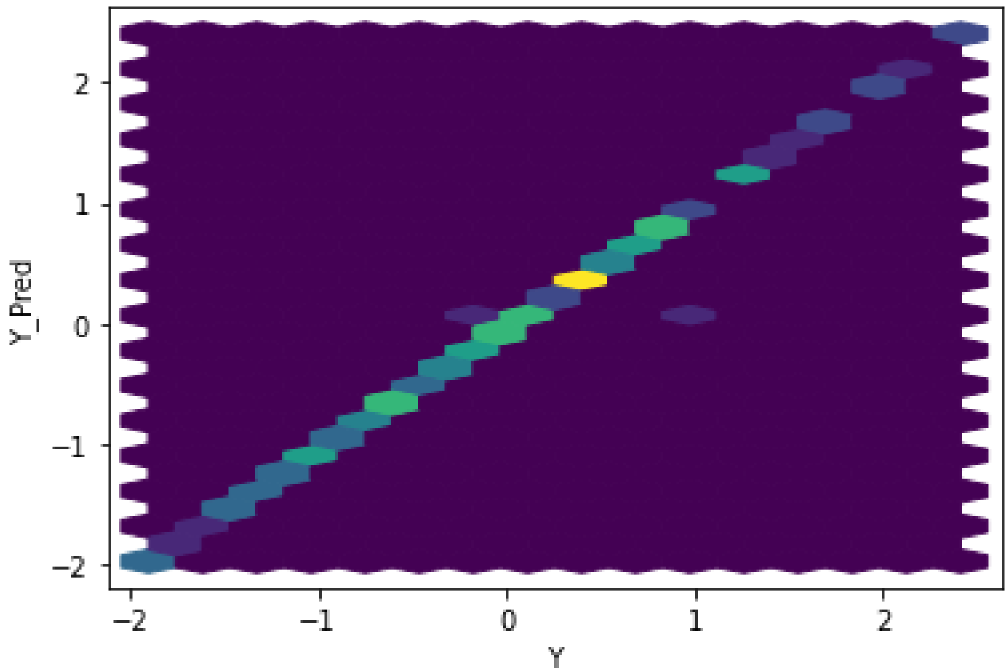
Y_Pred
Y

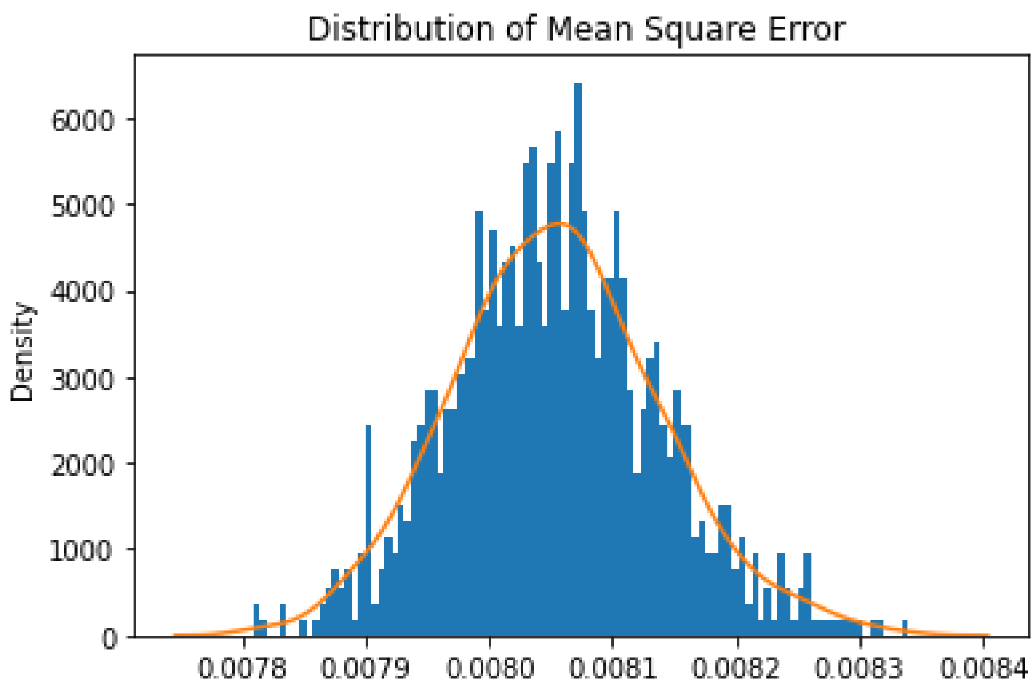


Y_Pred
Y

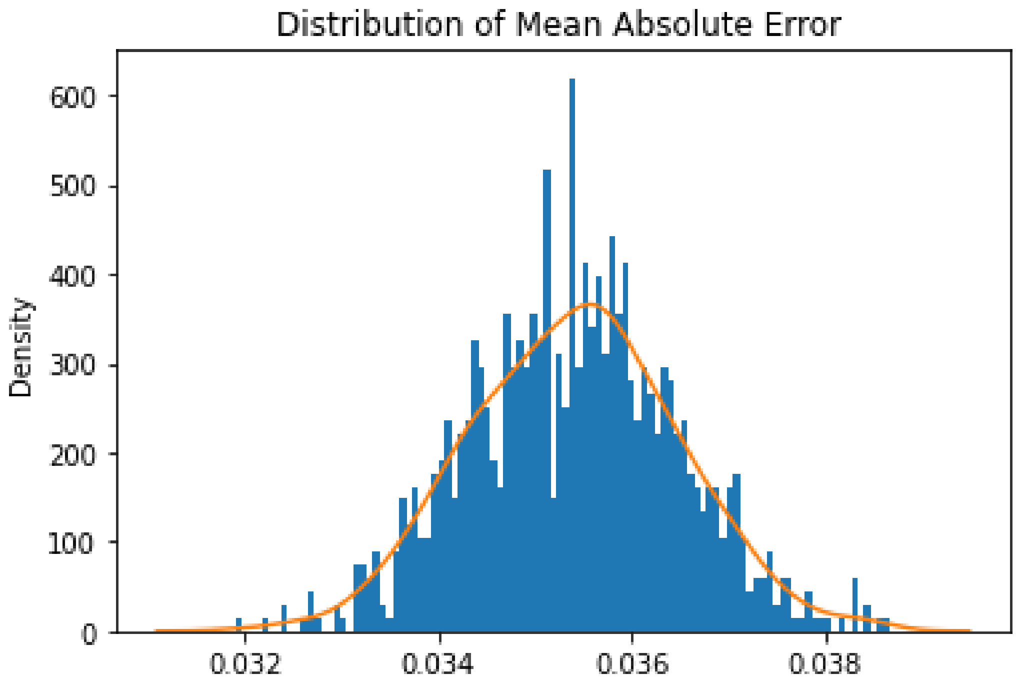

Y_Pred
Y
−2
−1
0
1
2


Y_Pred
Y
−2
−1
0
1
2

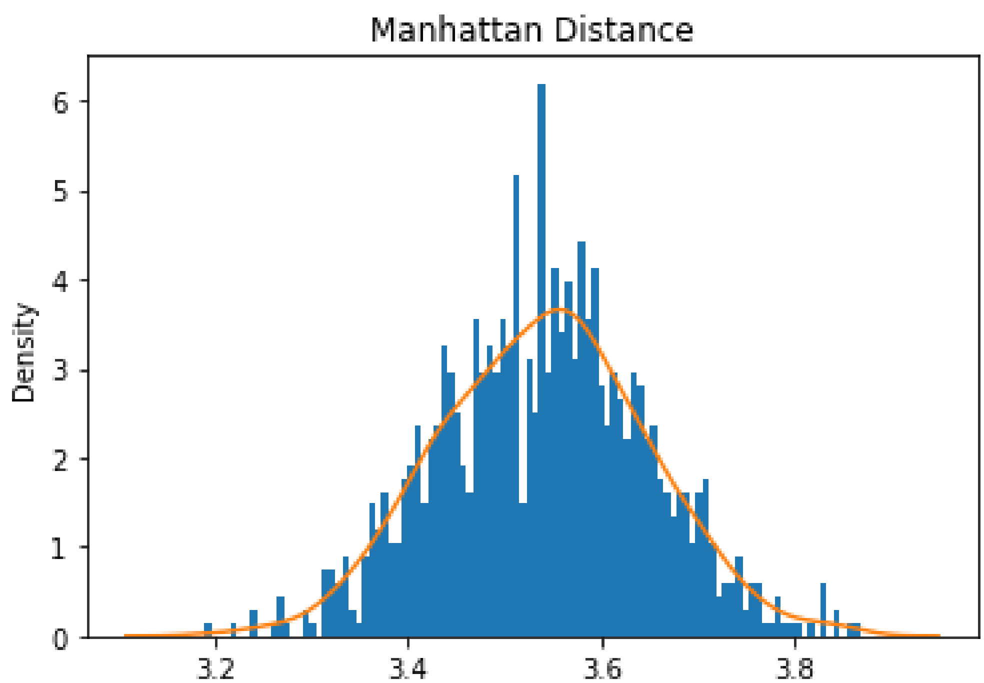




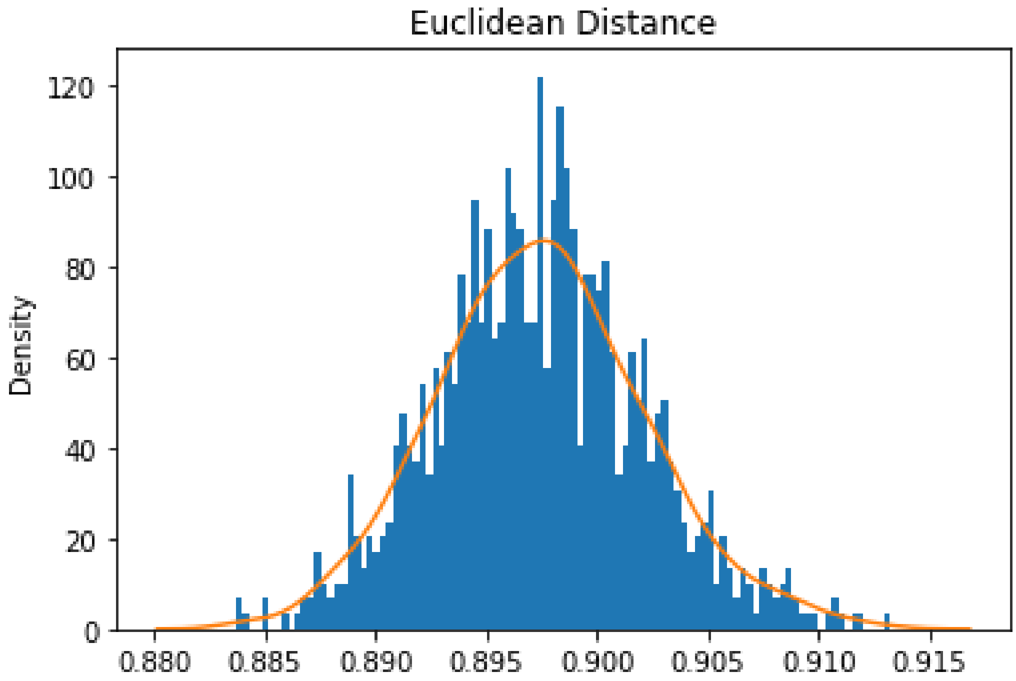
Mean Square Error: 0.008053967694310285




```
Mean Absolute Error: 0.03540318097885698
Mean Manhattan Distance: 3.5403180978856983
```

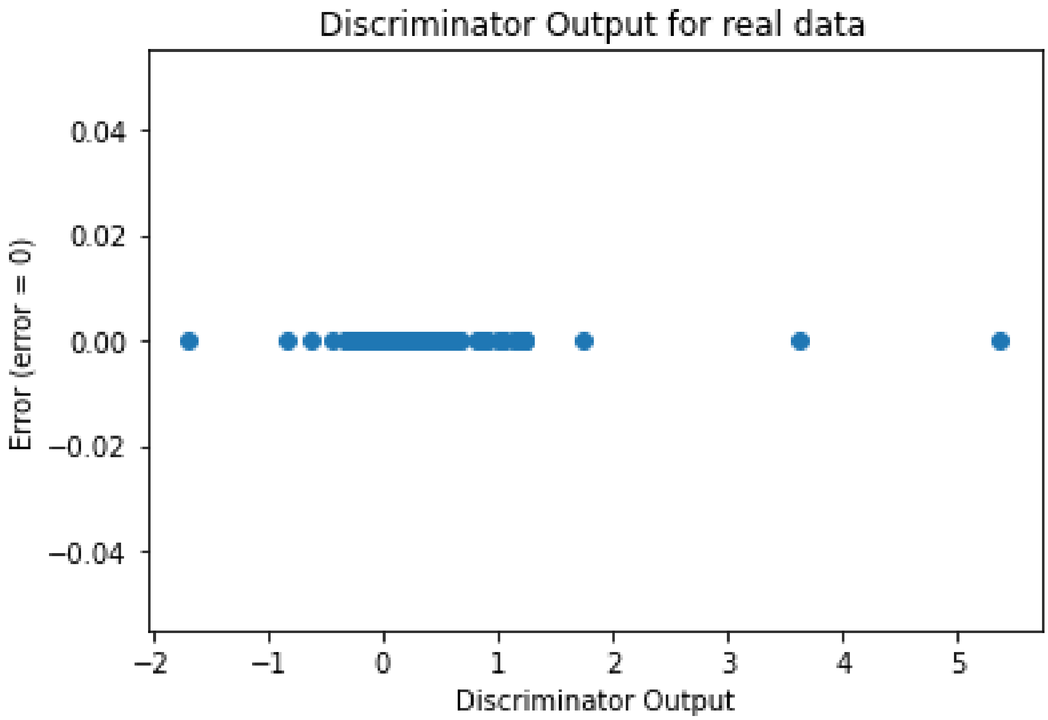


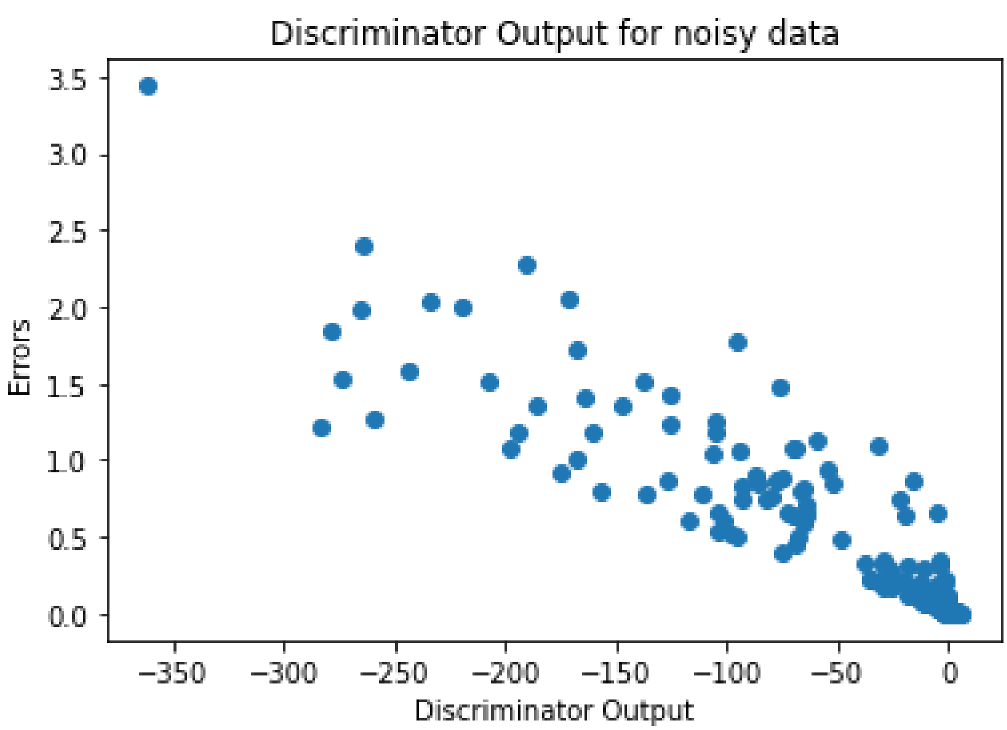
```
Mean Euclidean Distance: 0.8974266524814729
```



Sanity Checks

[20]:
```
sanityChecks.discProbVsError(real_dataset,disc,device)
```


Discriminator Output for real data
Error (error = 0)
0.04
0.02
0.00
−0.02
−0.04
−2
−1
0
1
2
3
4
5
Discriminator Output

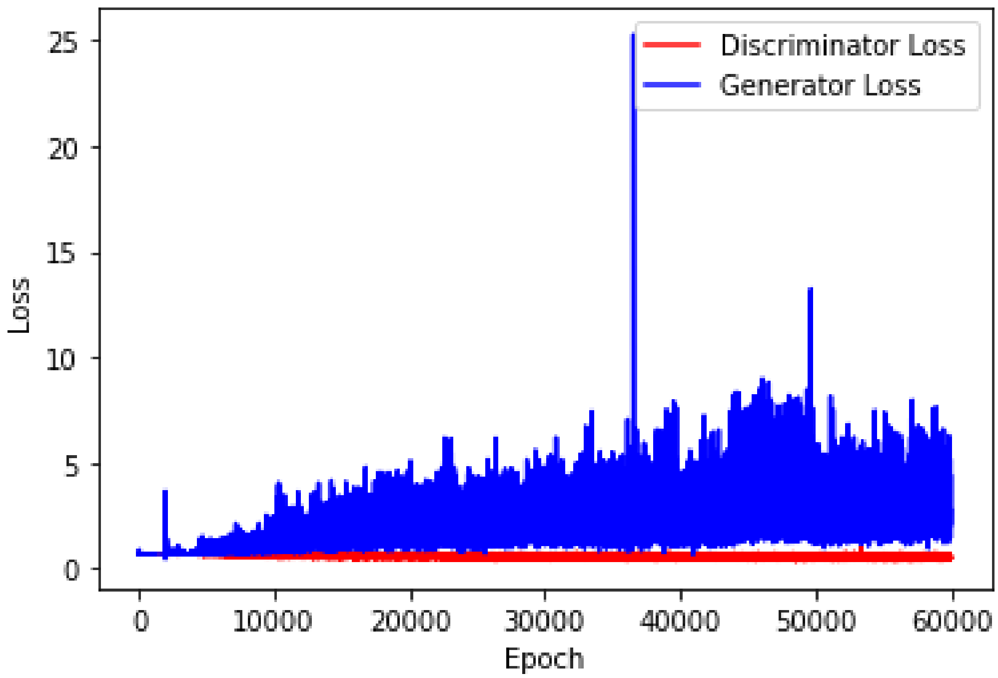

Discriminator Output for noisy data
Errors
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−350
−300
−250
−200
−150
−100
−50
0
Discriminator Output

Training GAN until mse of y_pred is > 0.1 or n_epochs < 30000

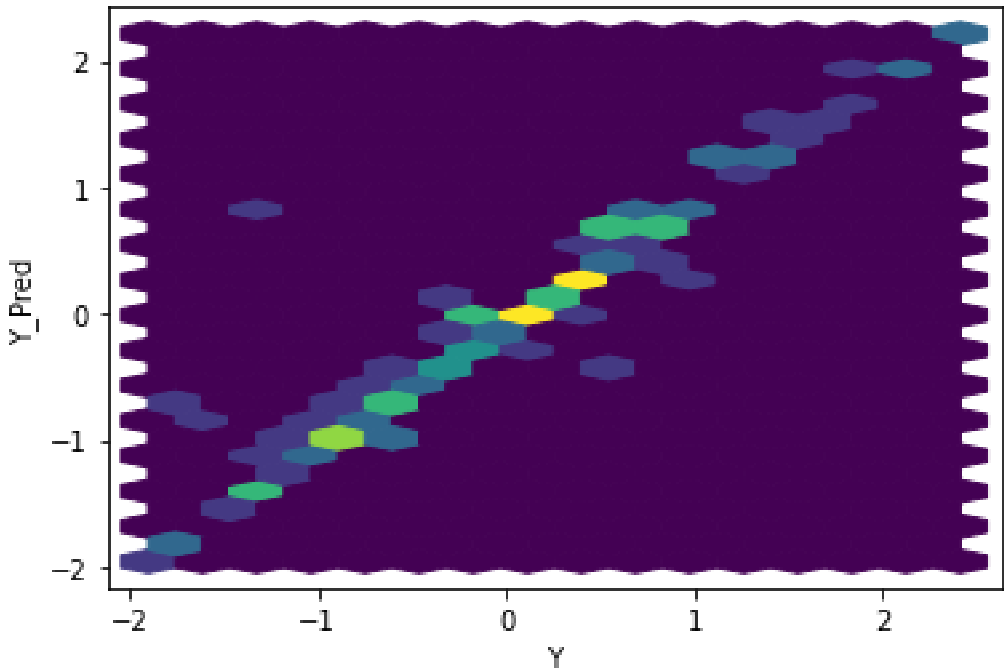
[21]:
```
gen = network.Generator(n_features+2)
disc = network.Discriminator(n_features+2)

criterion = torch.nn.BCEWithLogitsLoss()
gen_opt = torch.optim.Adam(gen.parameters(), lr=0.01, betas=(0.5, 0.999))
disc_opt = torch.optim.Adam(disc.parameters(), lr=0.01, betas=(0.5, 0.999))
```

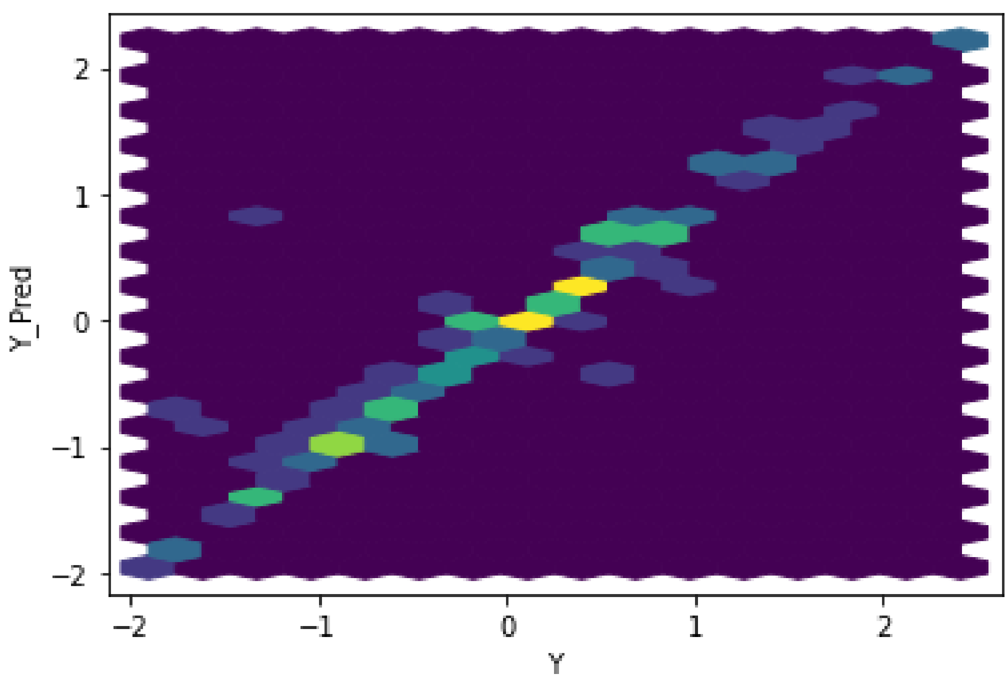
[22]:
```
ABC_train_test.
↪training_GAN_2(disc,gen,disc_opt,gen_opt,real_dataset,batch_size,␣
↪error,criterion,coeff,mean,variance,device)
```

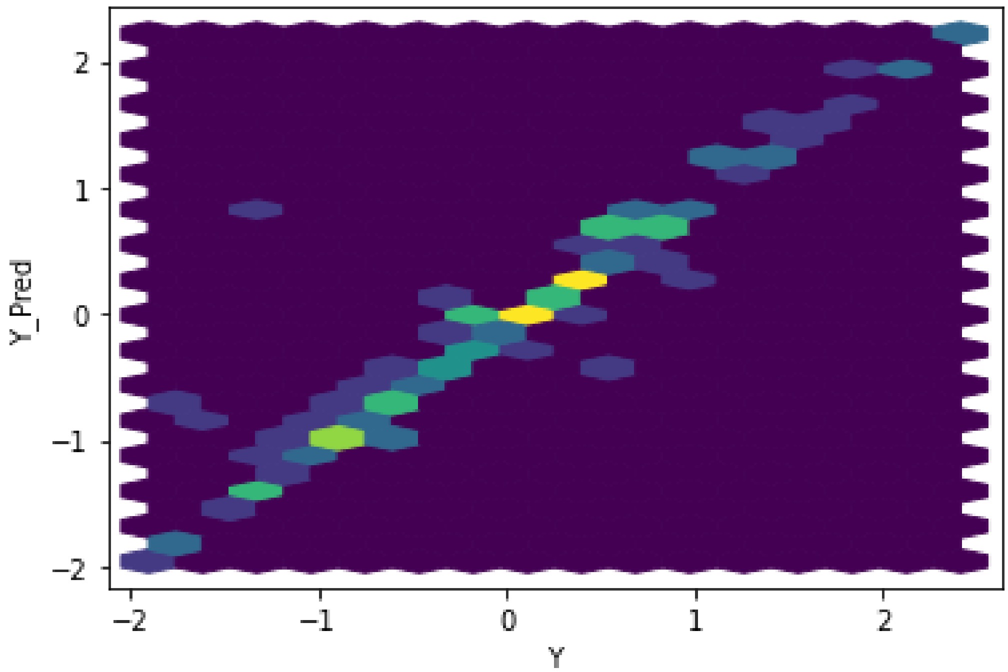
```
Number of epochs 30000
```

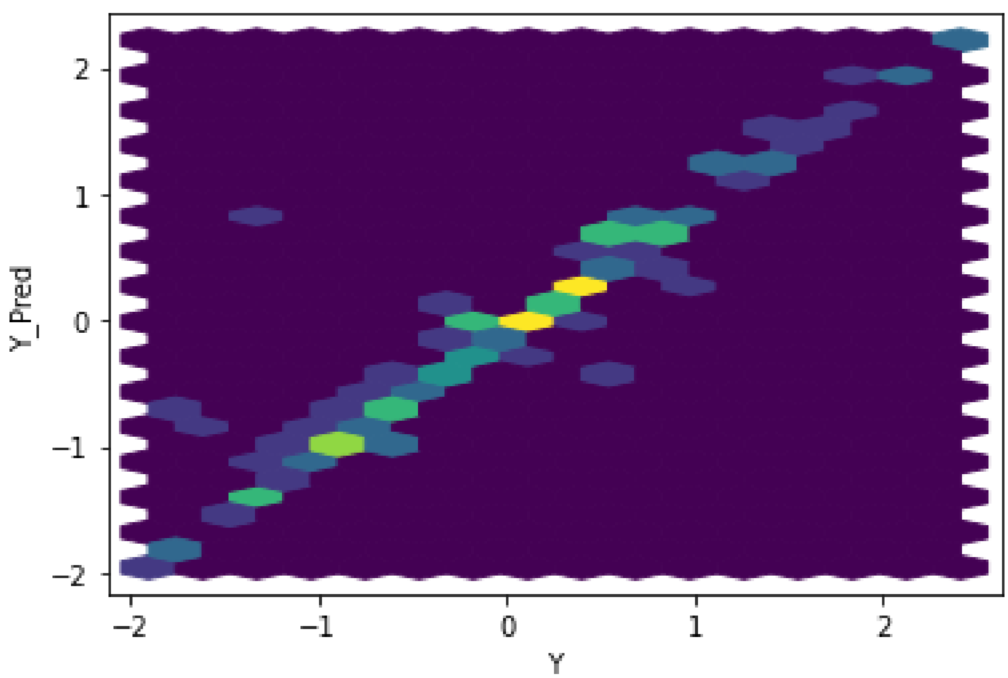


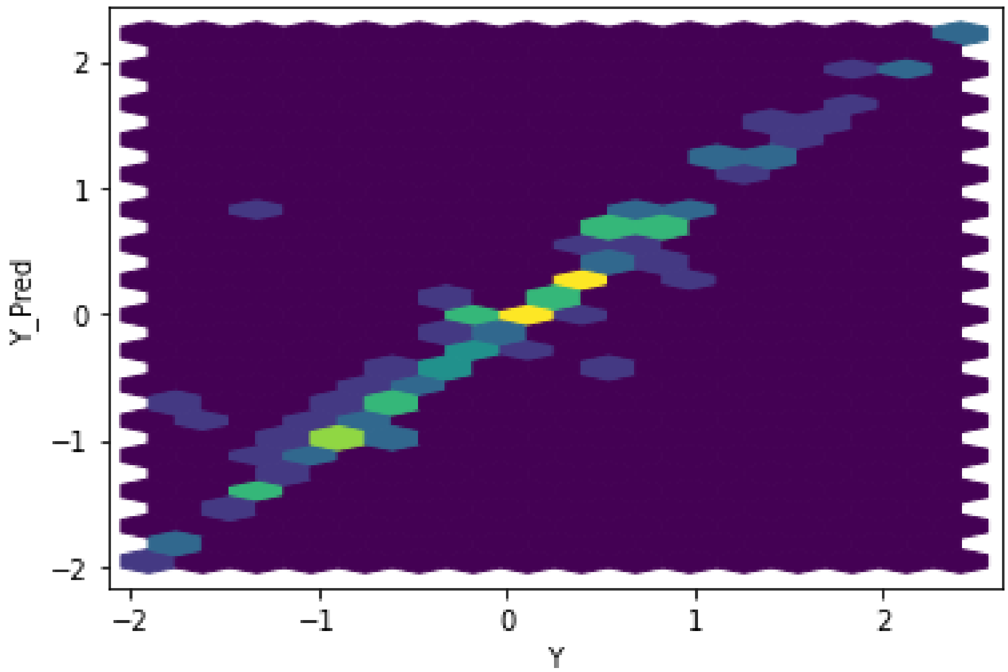
[23]:
```
ABC_train_test.test_generator(gen,real_dataset,coeff,mean,variance,device)
```

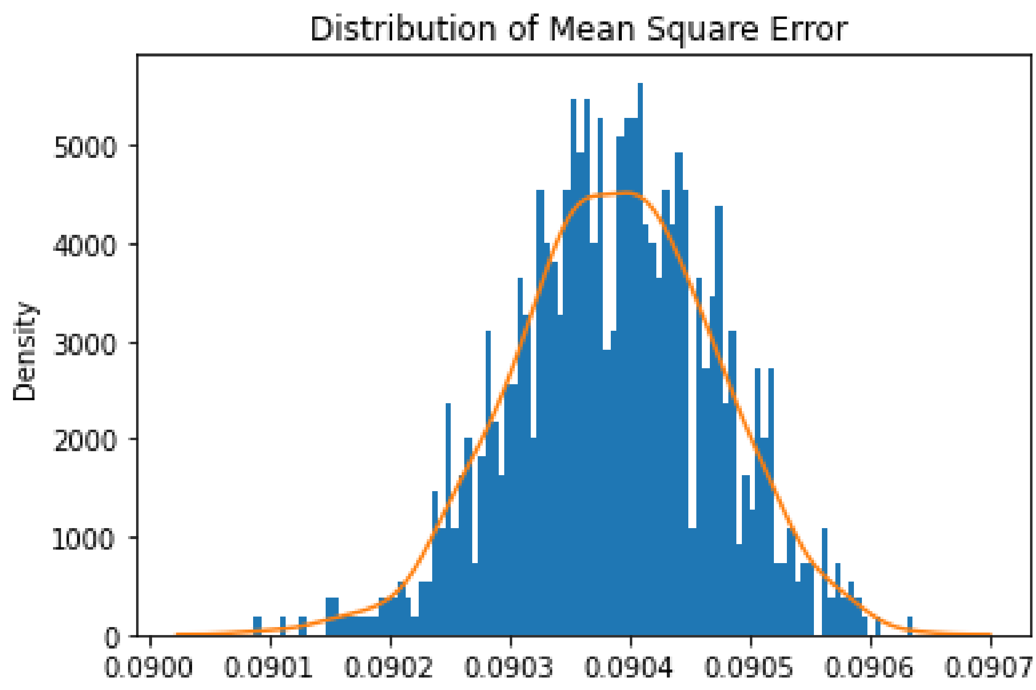

Y_Pred
Y


Y_Pred
Y

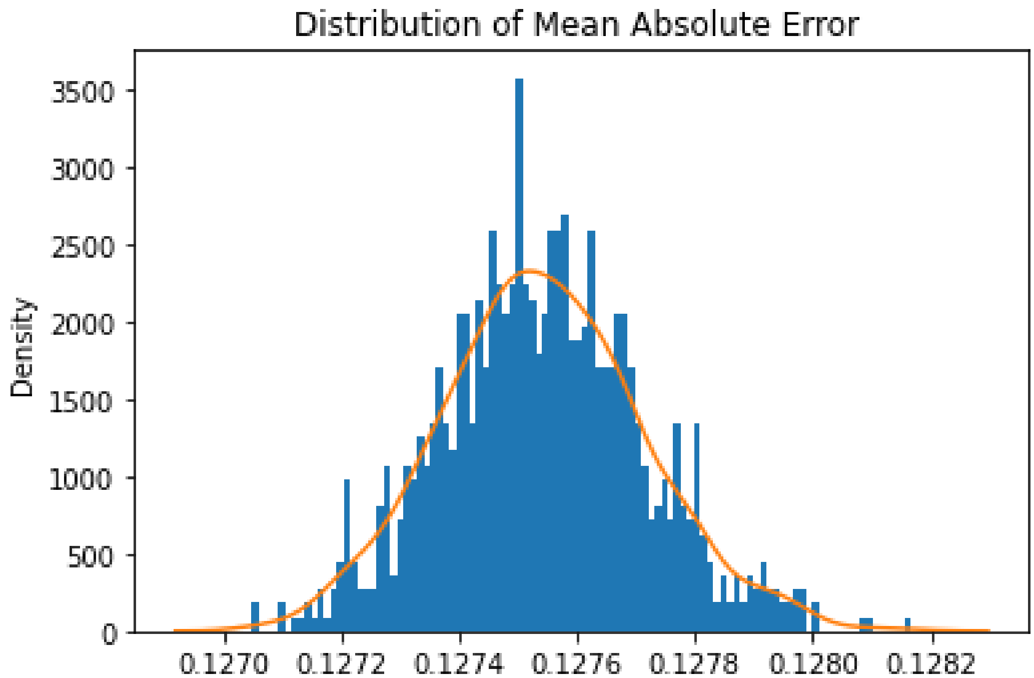

2
1
0
−1
−2
Y_Pred
−2
−1
0
1
2
Y


2
1
0
−1
−2
Y_Pred
−2
−1
0
1
2
Y

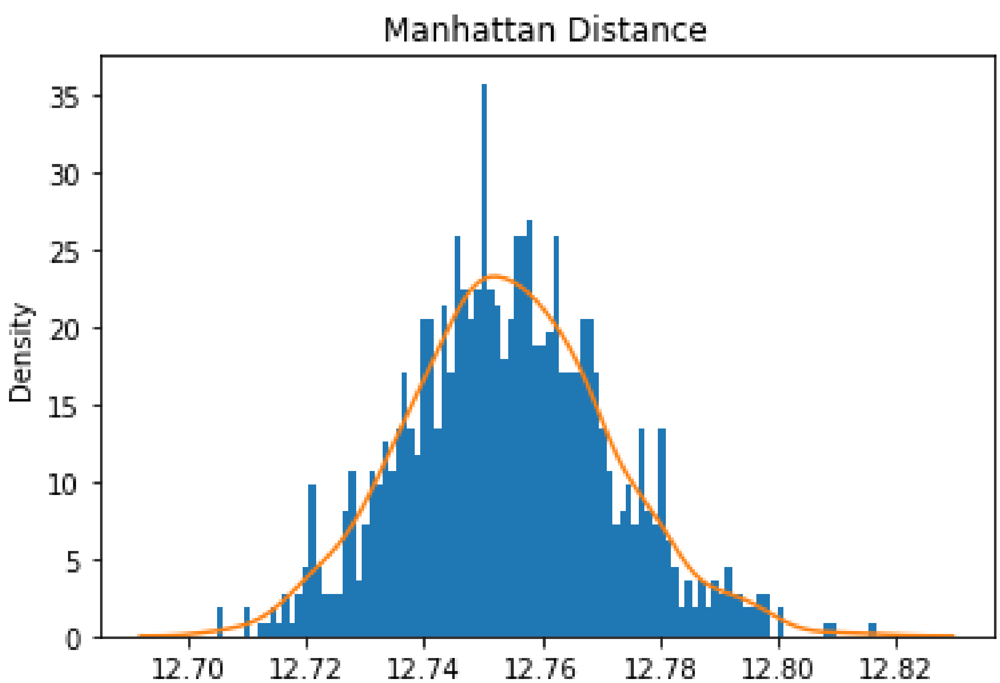




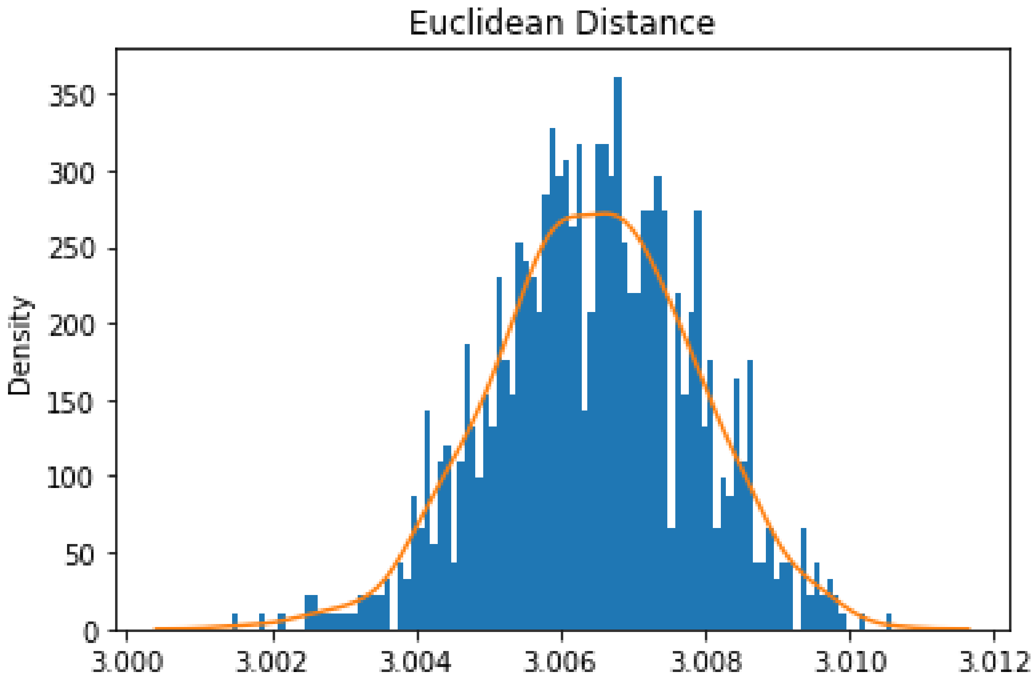

```
Mean Square Error: 0.09038765211938297
```




```
Mean Absolute Error: 0.12754193380184473
Mean Manhattan Distance: 12.754193380184471
```



```
Mean Euclidean Distance: 3.0064536017150707
```




```
[ ]:
```