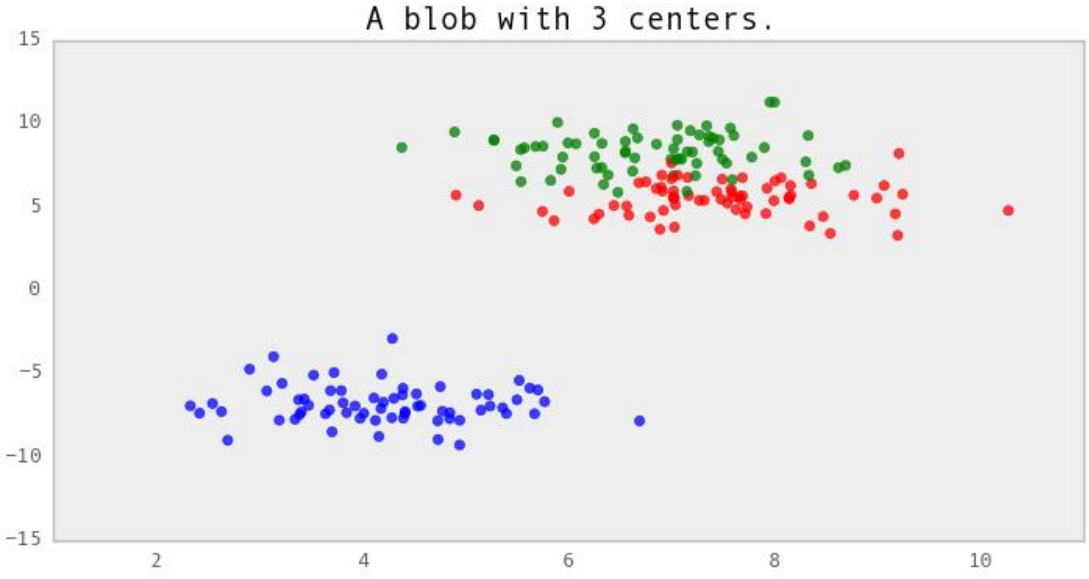
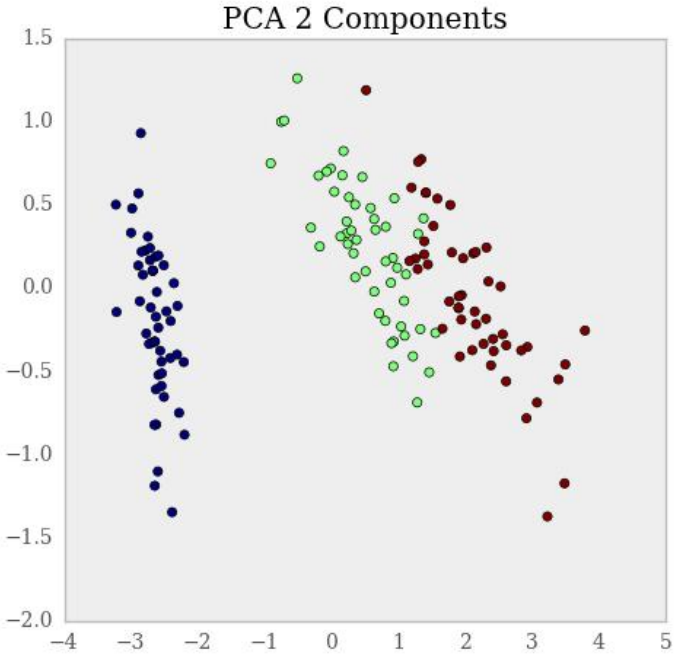
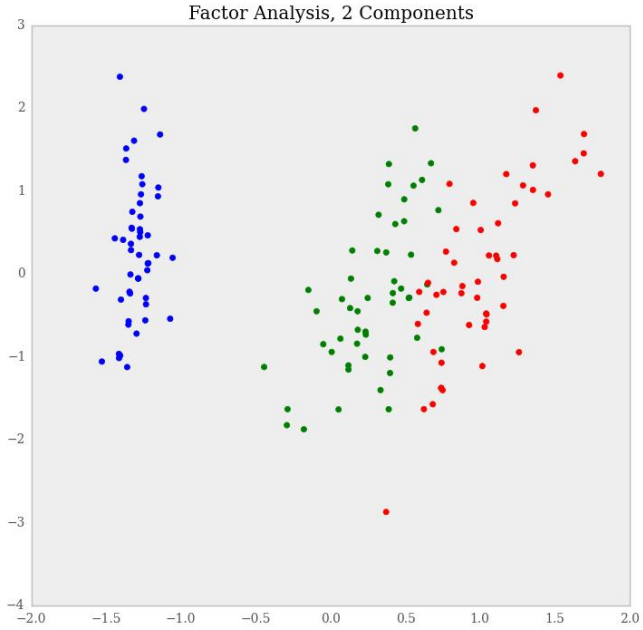
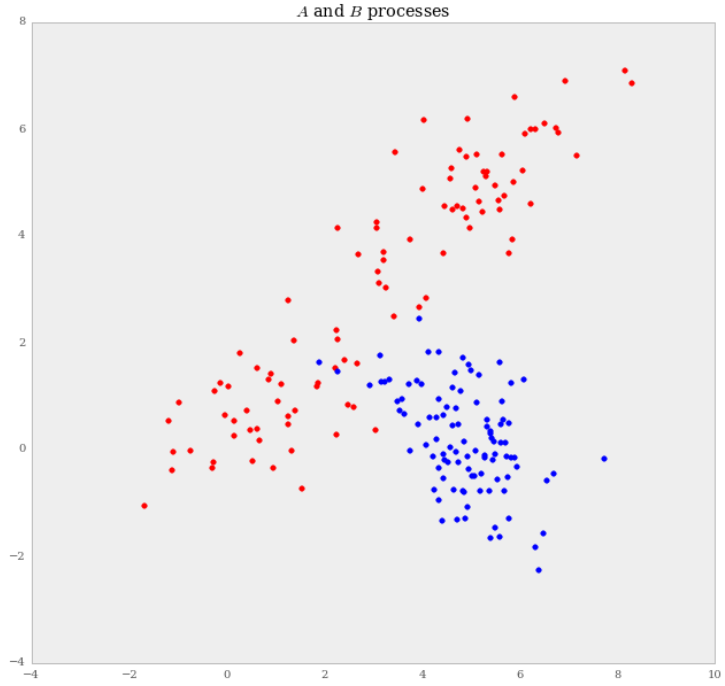
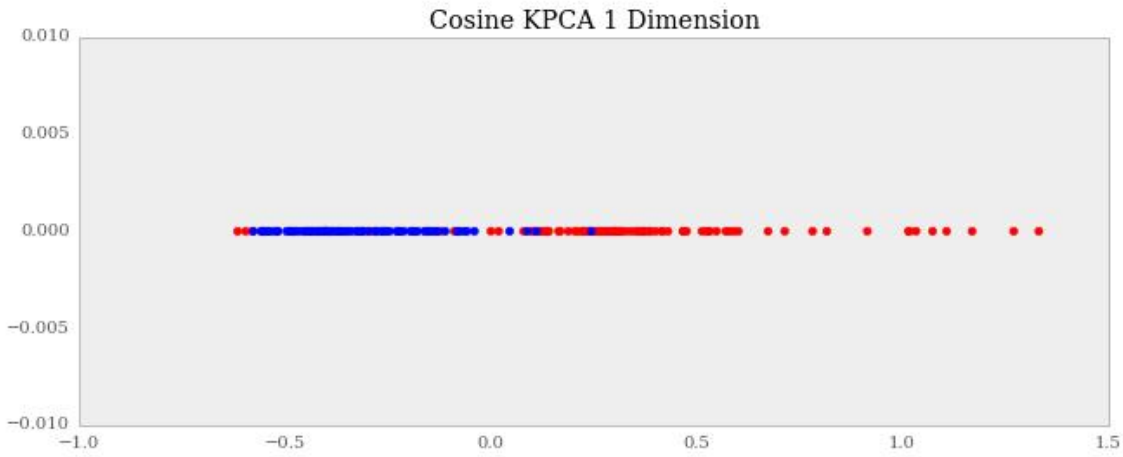
# Chapter 1

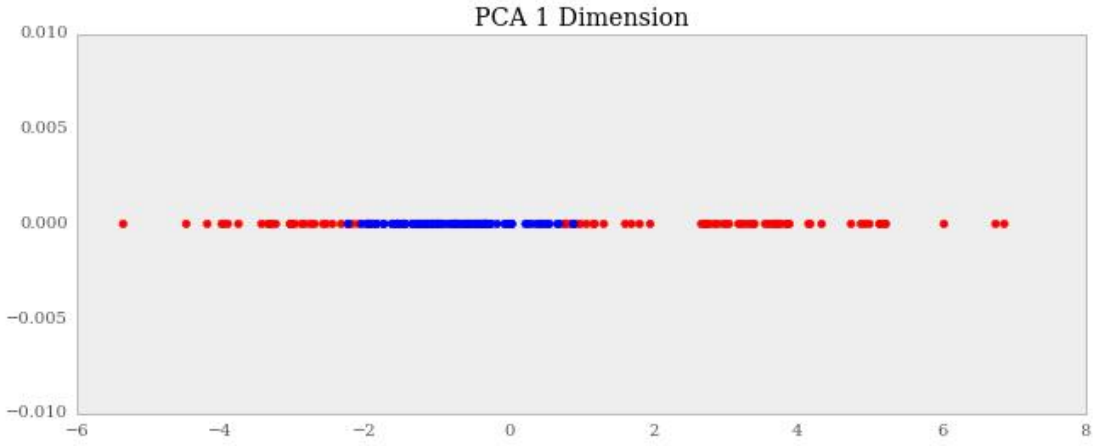

A blob with 3 centers.
15
10
5
0
−5
−10
−15
2
4
6
8
10

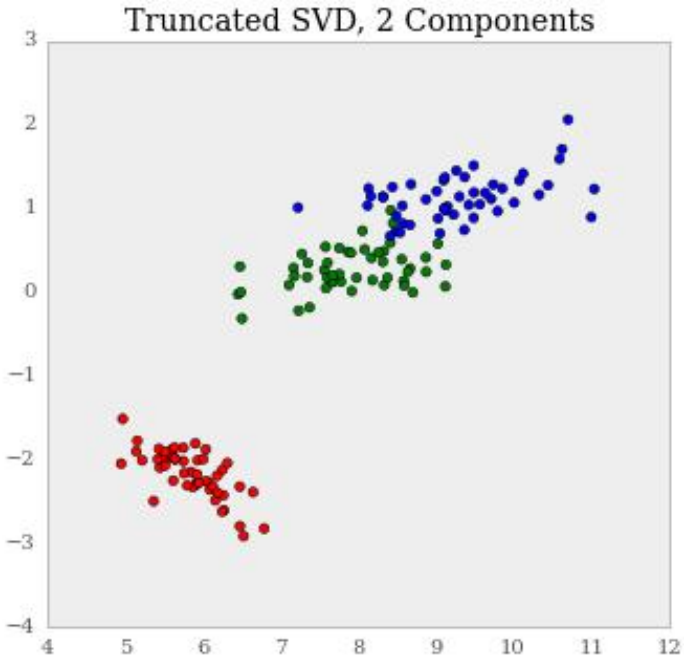

PCA 2 Components
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−4
−3
−2
−1
0
1
2
3
4
5

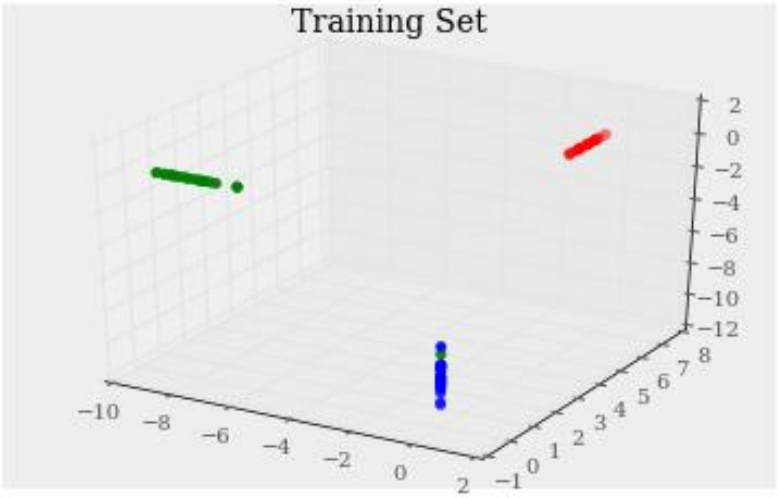

Factor Analysis, 2 Components
3
2
1
0
−1
−2
−3
−4
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0

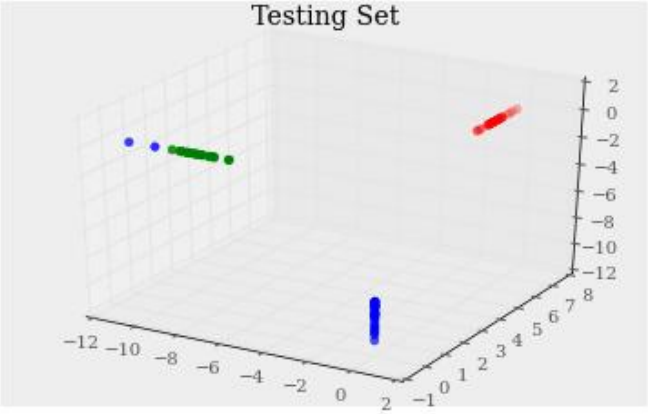

$A$ and $B$ processes
8
6
4
2
0
−2
−4
−4
−2
0
2
4
6
8
10

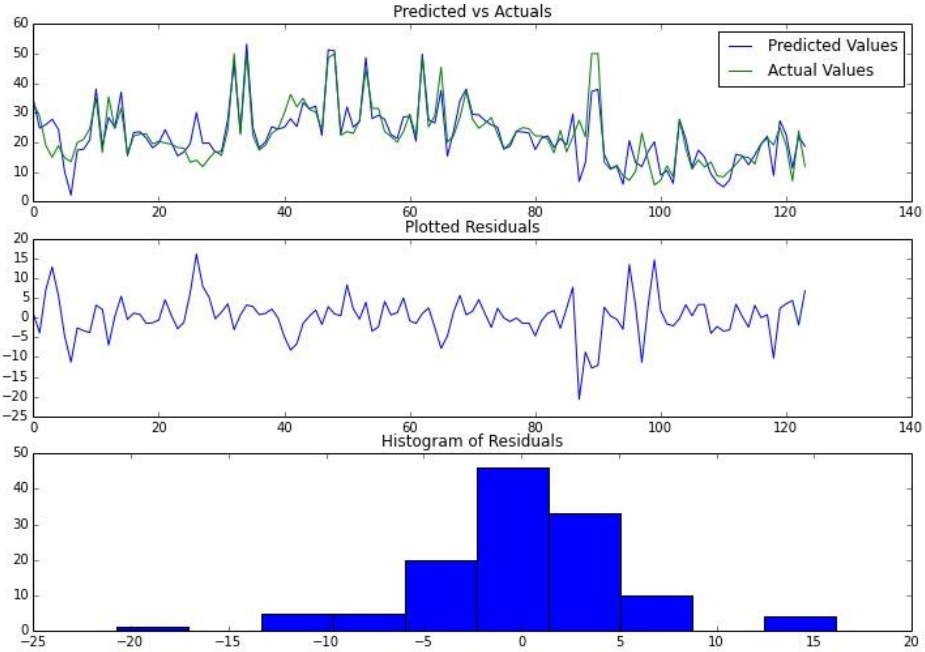

Cosine KPCA 1 Dimension
0.010
0.005
0.000
−0.005
−0.010
−1.0
−0.5
0.0
0.5
1.0
1.5

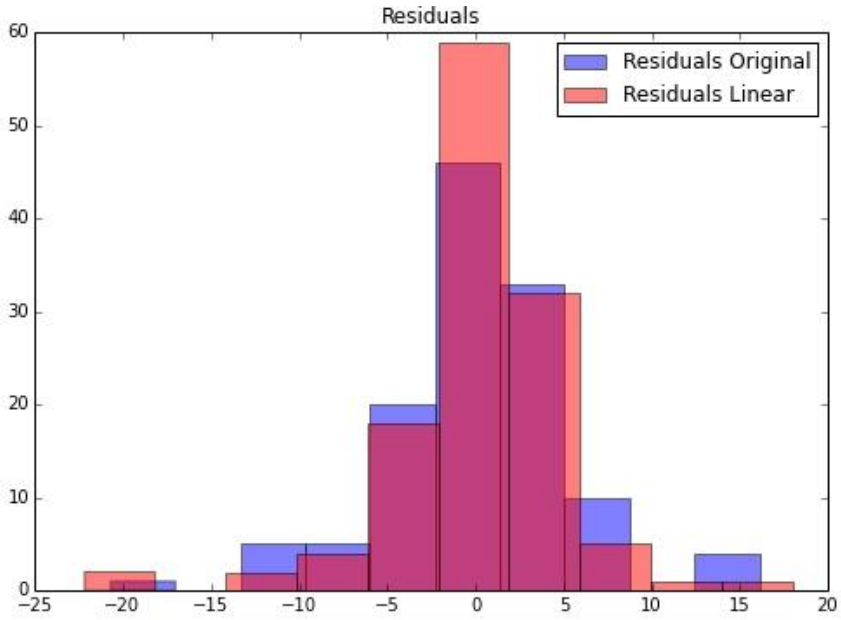

PCA 1 Dimension
0.010
0.005
0.000
−0.005
−0.010
−6
−4
−2
0
2
4
6
8

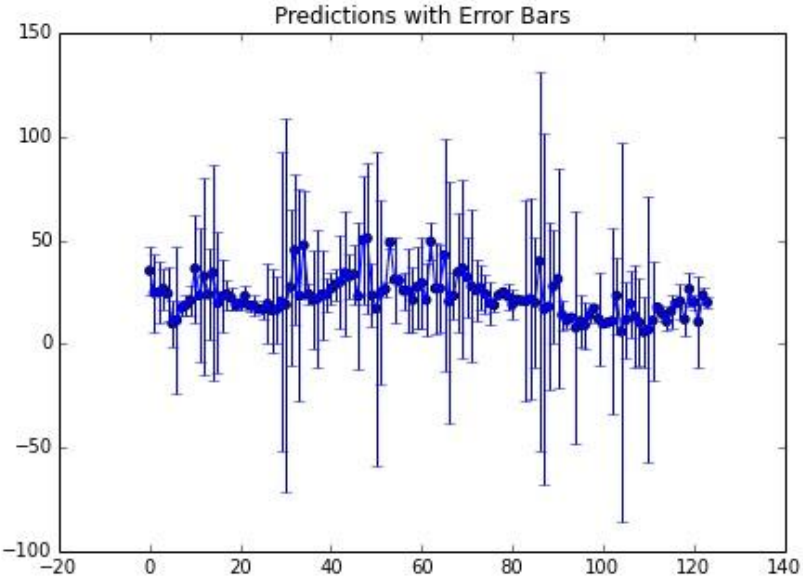

Truncated SVD, 2 Components
3
2
1
0
−1
−2
−3
−4
4
5
6
7
8
9
10
11
12

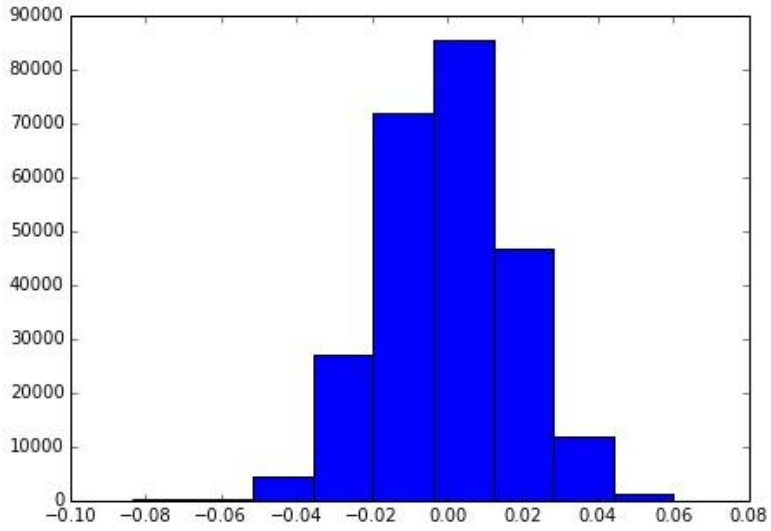

Training Set
2
0
−2
−4
−6
−8
−10
−12
−10
−8
−6
−4
−2
0
2
−1
0
1
2
3
4
5
6
7
8


Testing Set

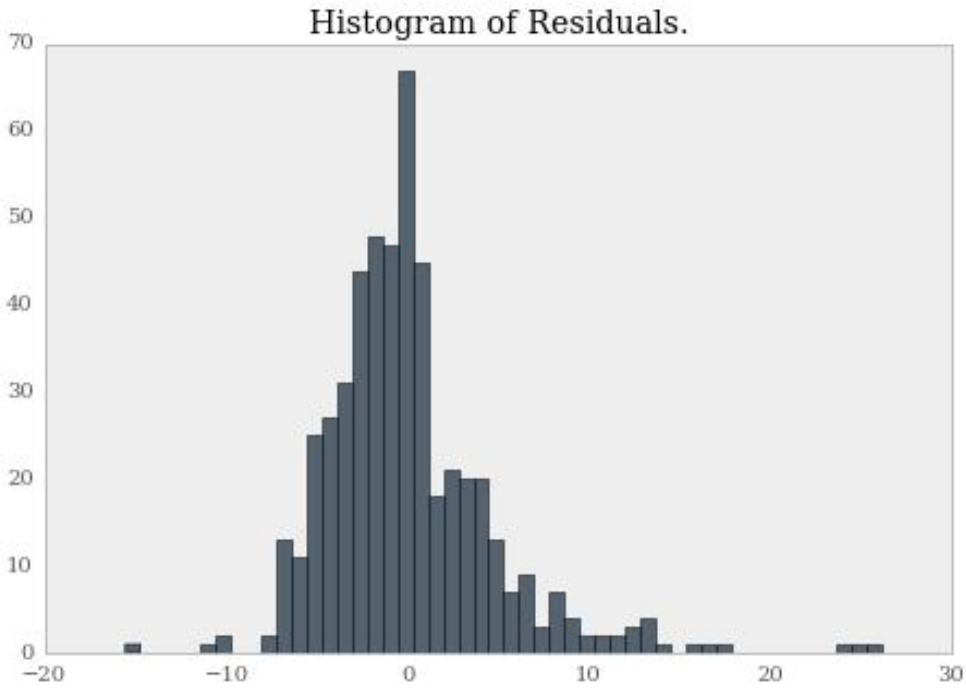

Predicted vs Actuals
Predicted Values
Actual Values
Plotted Residuals
Histogram of Residuals

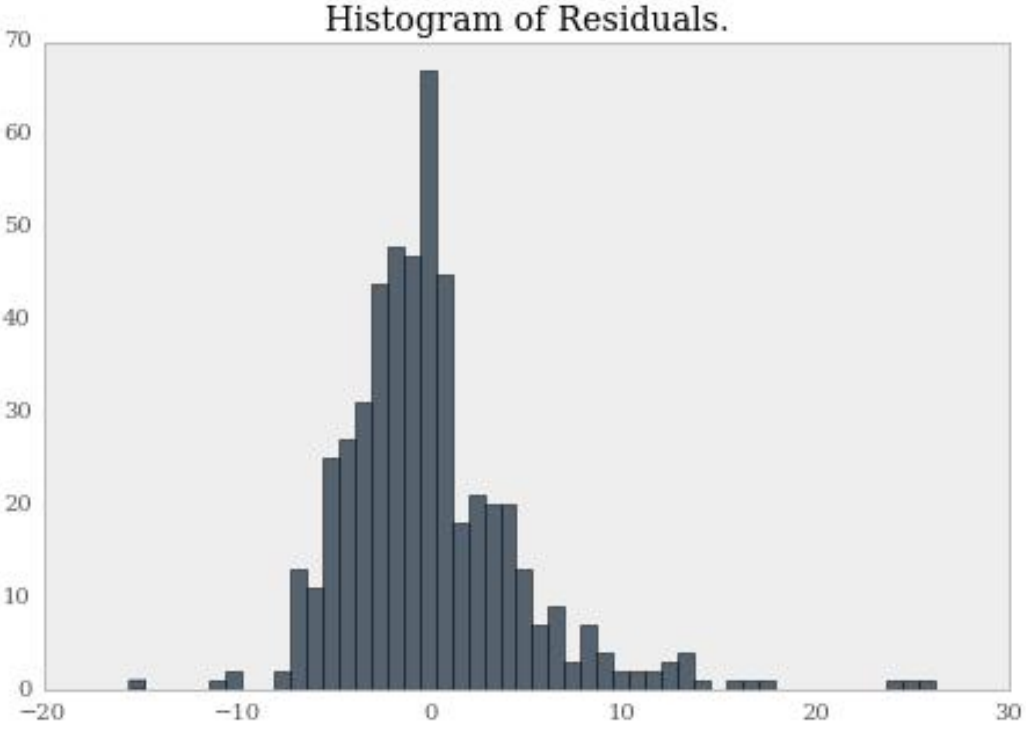

Residuals
Residuals Original
Residuals Linear
60
50
40
30
20
10
0
−25
−20
−15
−10
−5
0
5
10
15
20

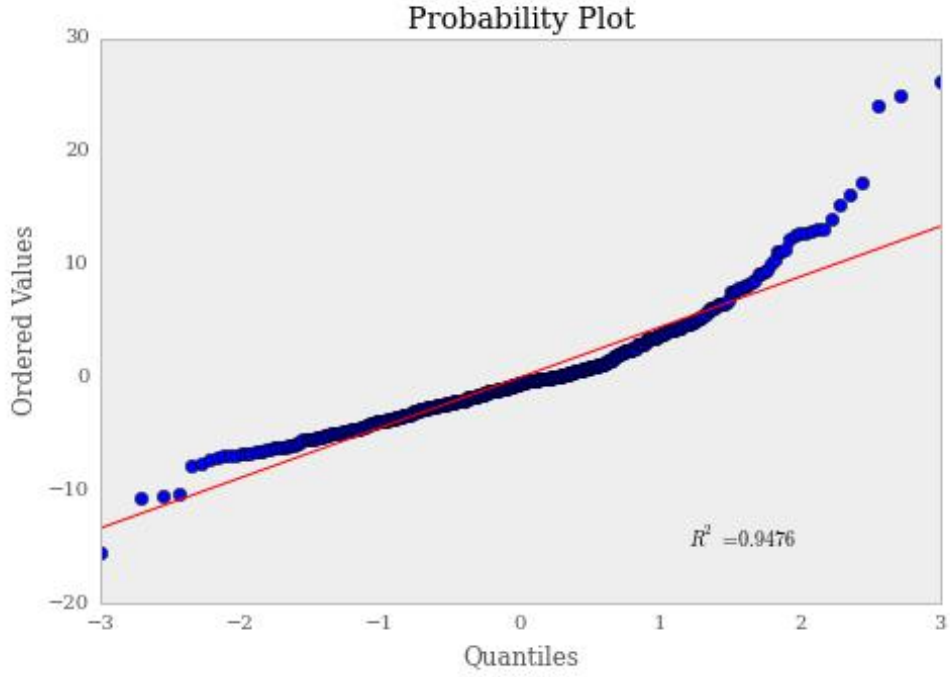

Predictions with Error Bars
150
100
50
0
−50
−100
−20
0
20
40
60
80
100
120
140

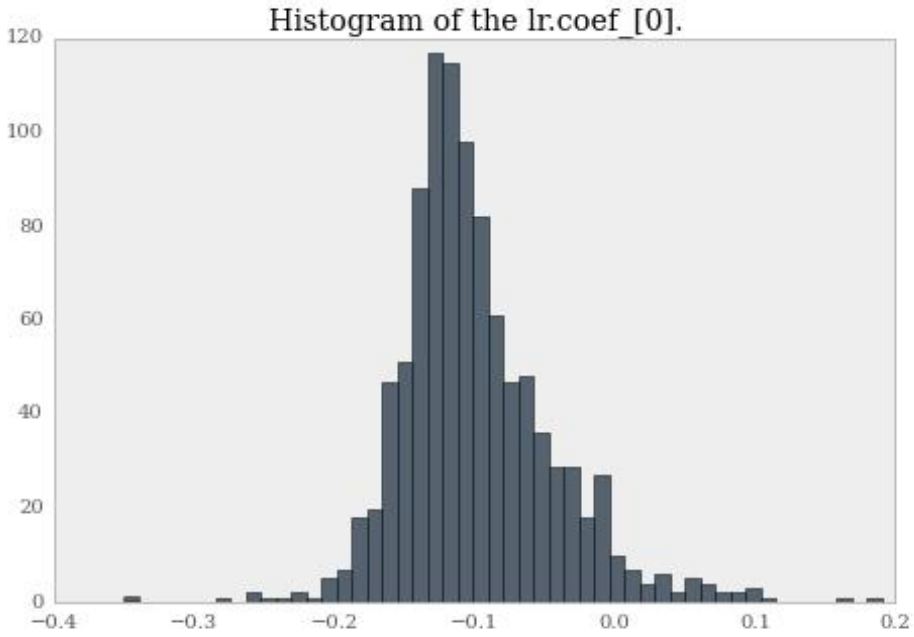

90000
80000
70000
60000
50000
40000
30000
20000
10000
0
−0.10
−0.08
−0.06
−0.04
−0.02
0.00
0.02
0.04
0.06
0.08

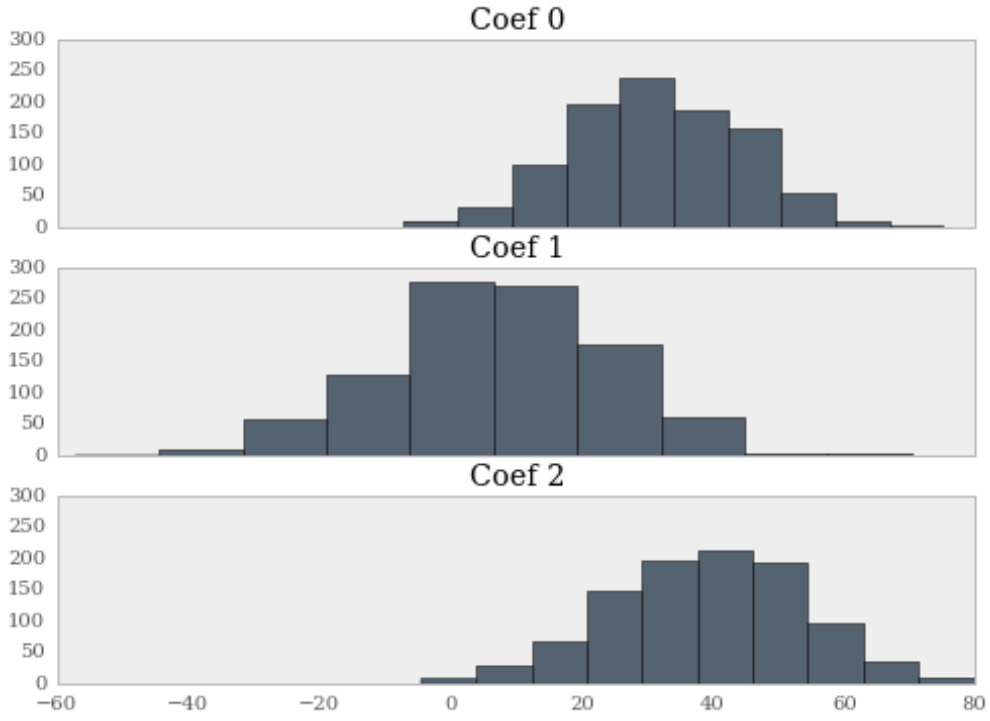
## Chapter 2

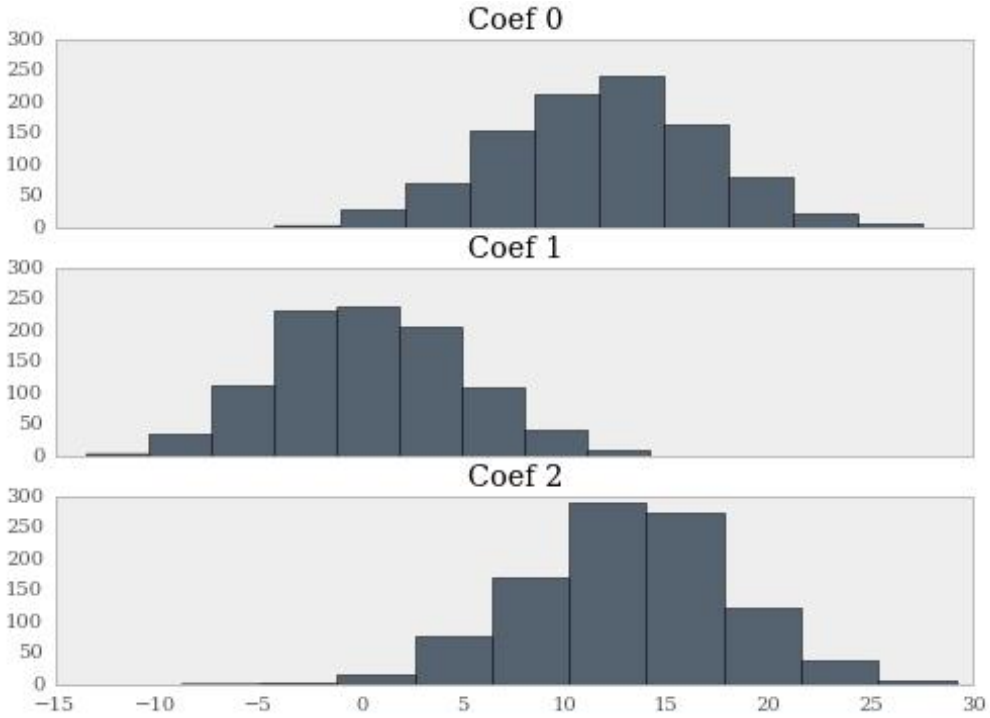


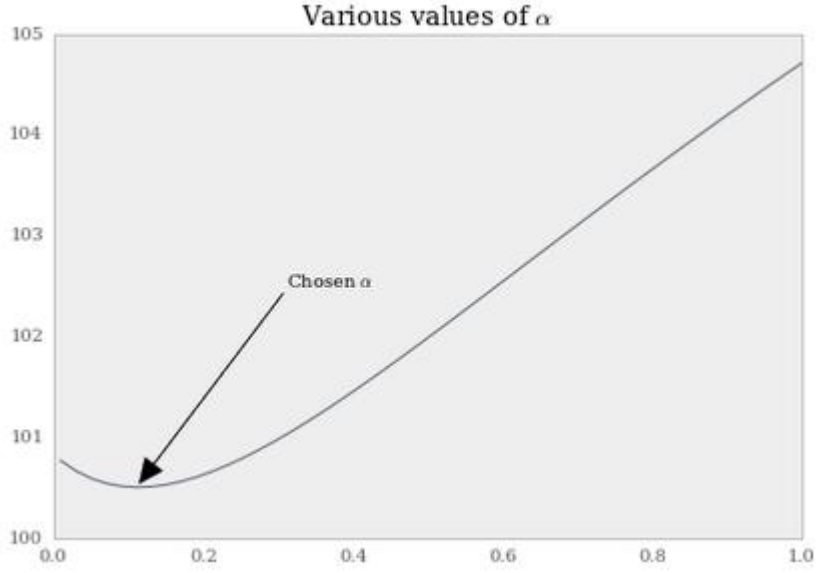

Histogram of Residuals.
70
60
50
40
30
20
10
0
−20
−10
0
10
20
30

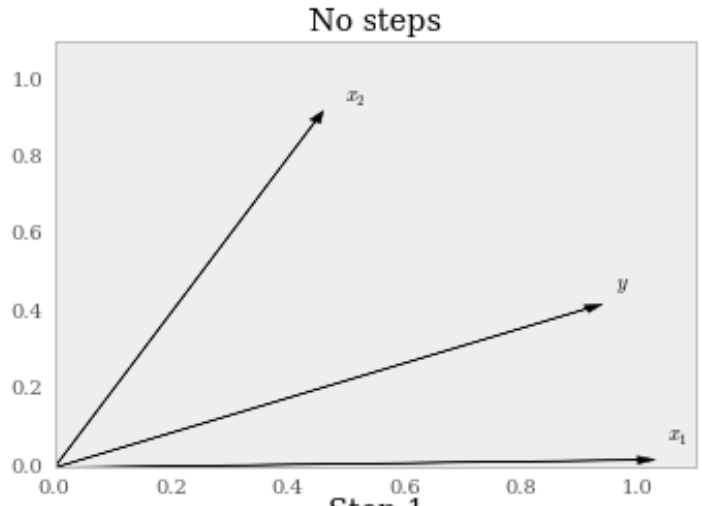

Probability Plot
Ordered Values
Quantiles
$R^2$ =0.9476
30
20
10
0
−10
−20
−3
−2
−1
0
1
2
3

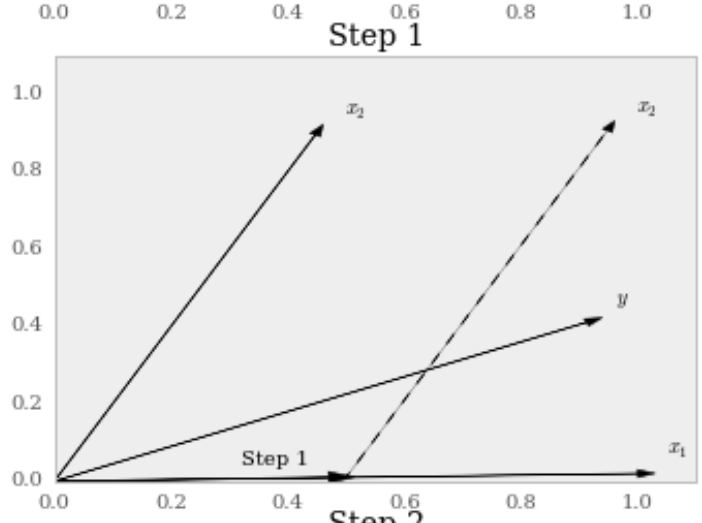

Histogram of the lr.coef_[0].
120
100
80
60
40
20
0
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2

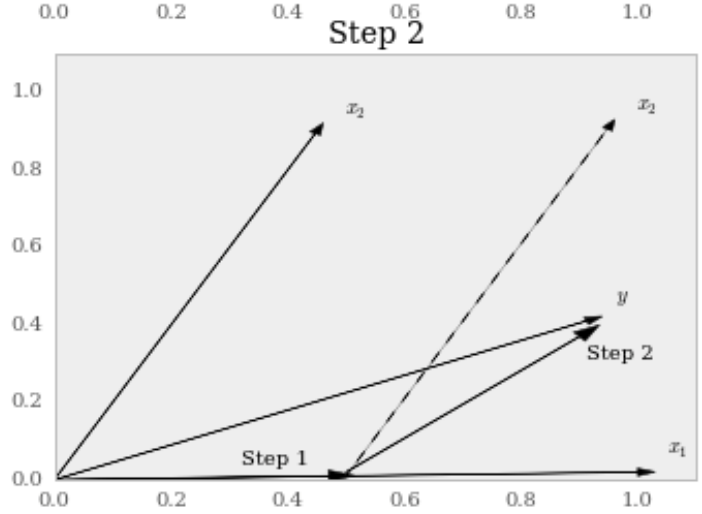

Coef 0
300
250
200
150
100
50
0
Coef 1
300
250
200
150
100
50
0
Coef 2
300
250
200
150
100
50
0
−60
−40
−20
0
20
40
60
80

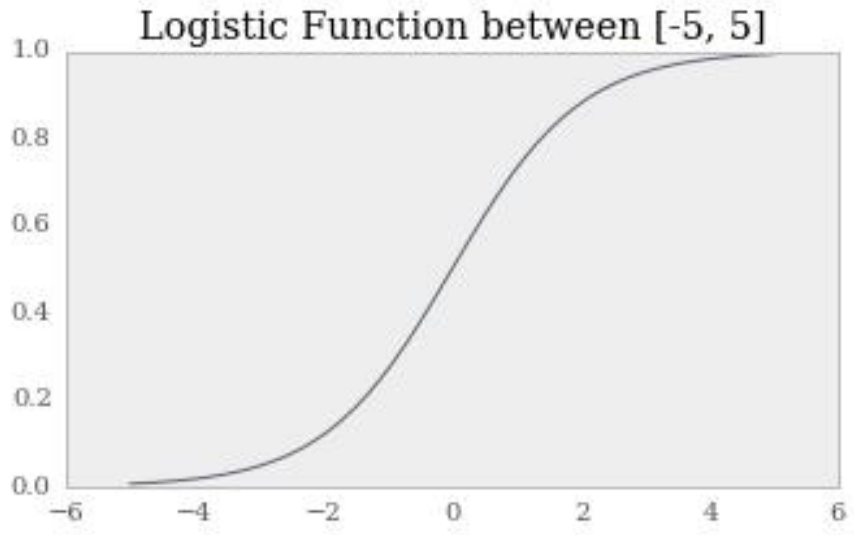

Coef 0
Coef 1
Coef 2
300
250
200
150
100
50
0
−15
−10
−5
0
5
10
15
20
25
30

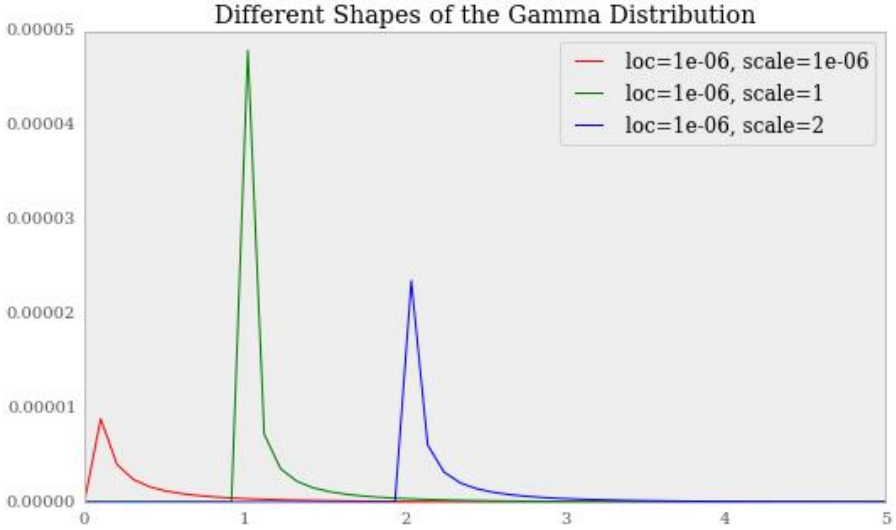

Various values of α
Chosen α
105
104
103
102
101
100
0.0
0.2
0.4
0.6
0.8
1.0

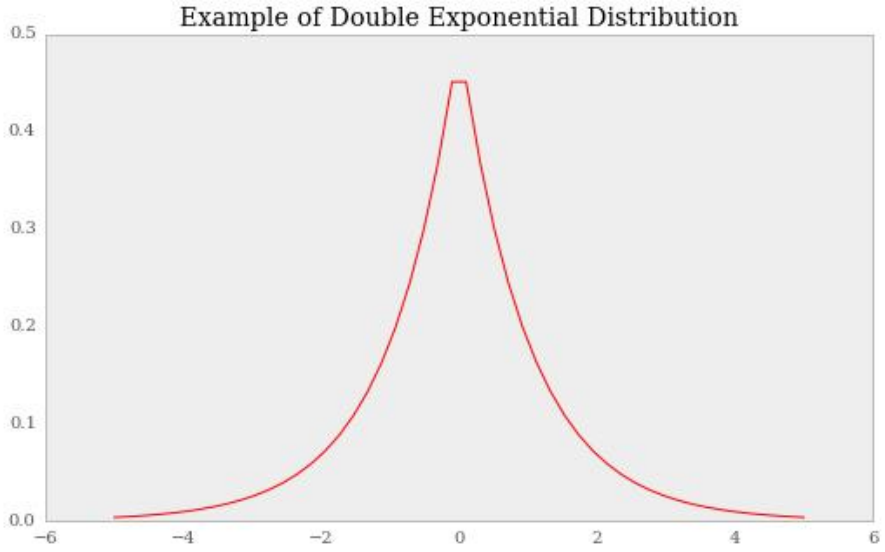

No steps
1.0
0.8
0.6
0.4
0.2
0.0
$x_2$
$y$
$x_1$
0.0
0.2
0.4
0.6
0.8
1.0

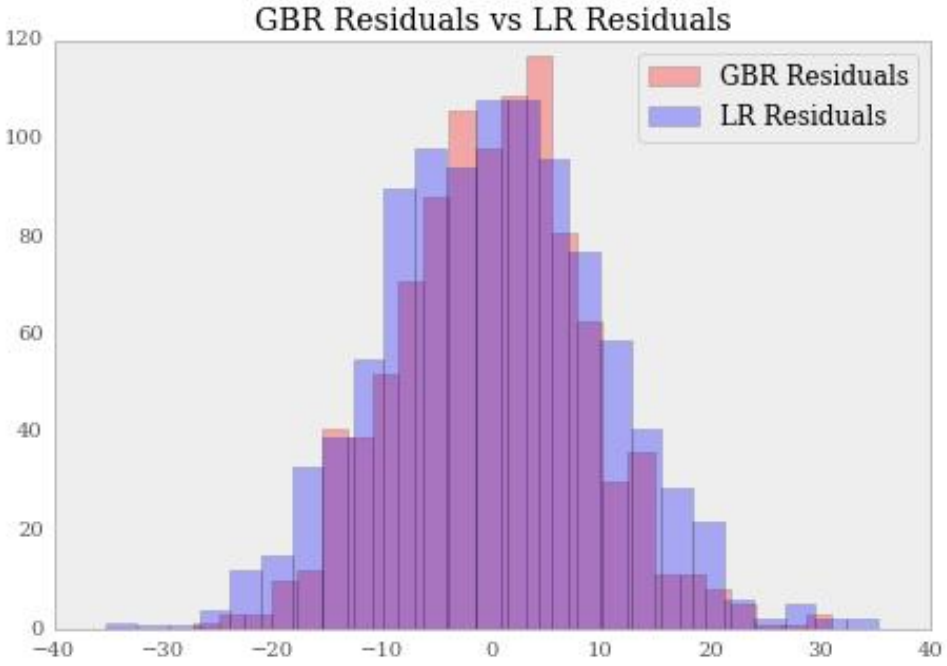

Step 1
1.0
0.8
0.6
0.4
0.2
0.0
$x_2$
$x_2$
$y$
Step 1
$x_1$
0.0
0.2
0.4
0.6
0.8
1.0

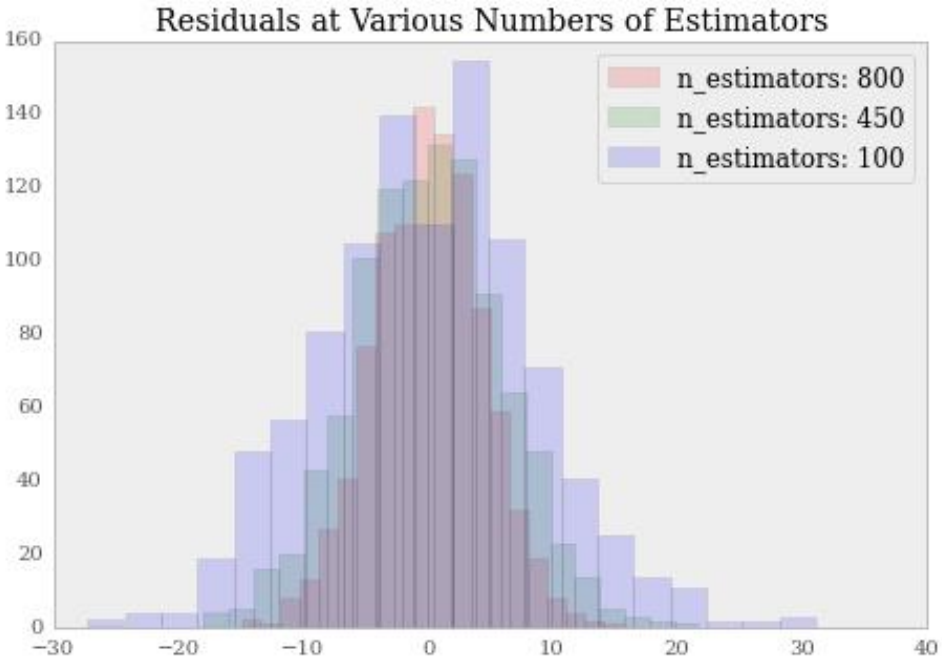

Step 2
1.0
0.8
0.6
0.4
0.2
0.0
$x_2$
$x_2$
$y$
Step 2
Step 1
$x_1$
0.0
0.2
0.4
0.6
0.8
1.0

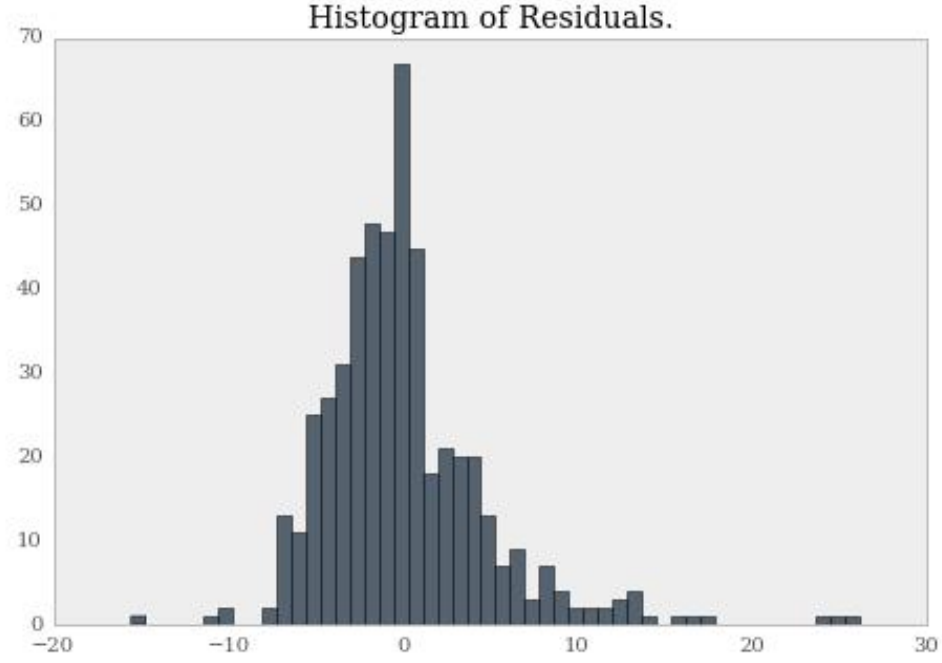

Logistic Function between [-5, 5]
1.0
0.8
0.6
0.4
0.2
0.0
−6
−4
−2
0
2
4
6

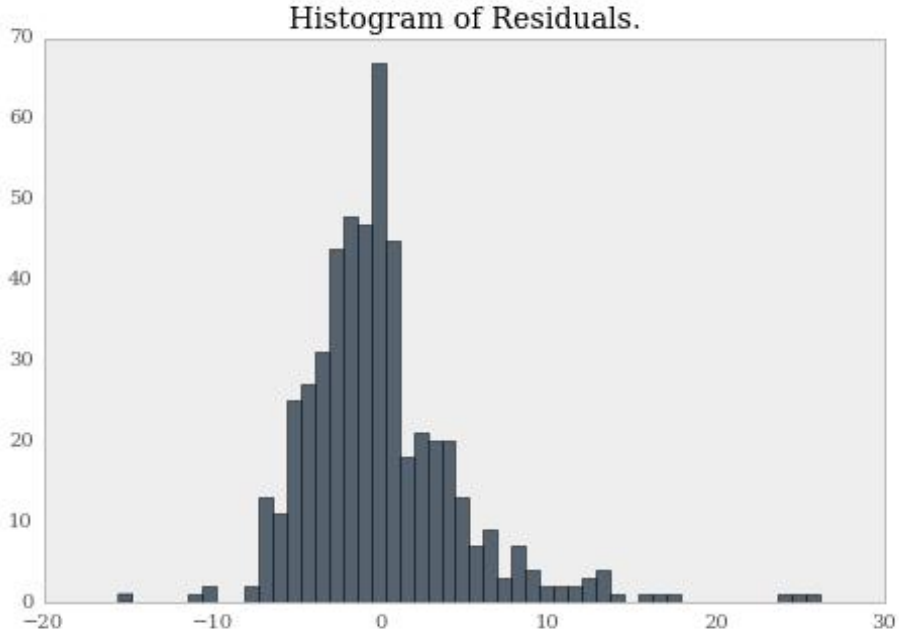

Different Shapes of the Gamma Distribution
loc=1e-06, scale=1e-06
loc=1e-06, scale=1
loc=1e-06, scale=2
0.00005
0.00004
0.00003
0.00002
0.00001
0.00000
0
1
2
3
4
5

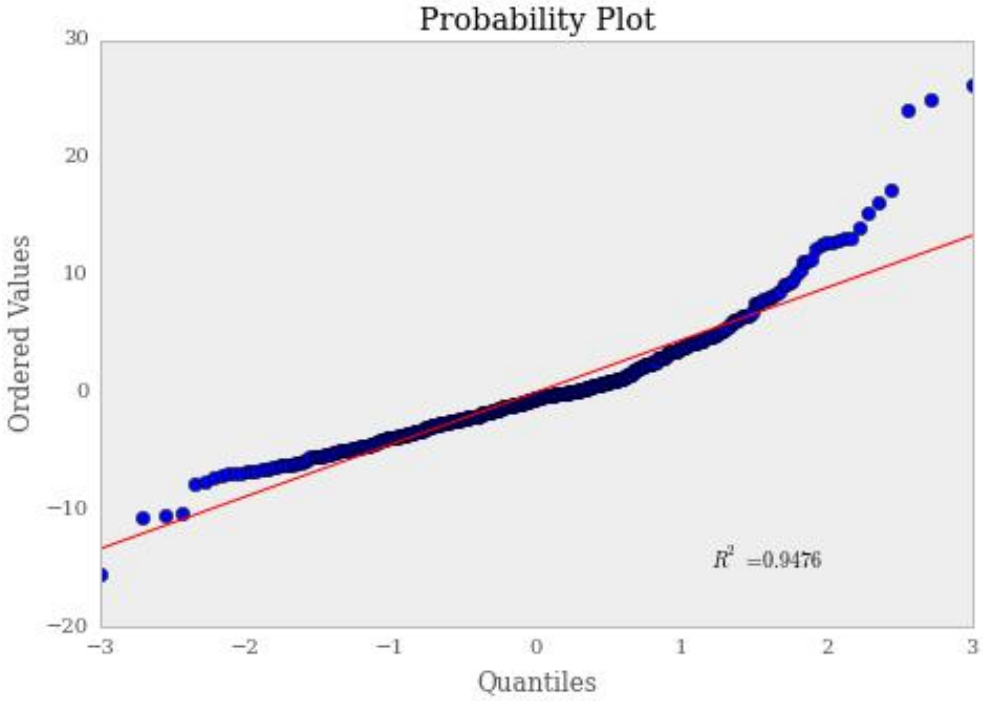

Example of Double Exponential Distribution
0.5
0.4
0.3
0.2
0.1
0.0
−6
−4
−2
0
2
4
6

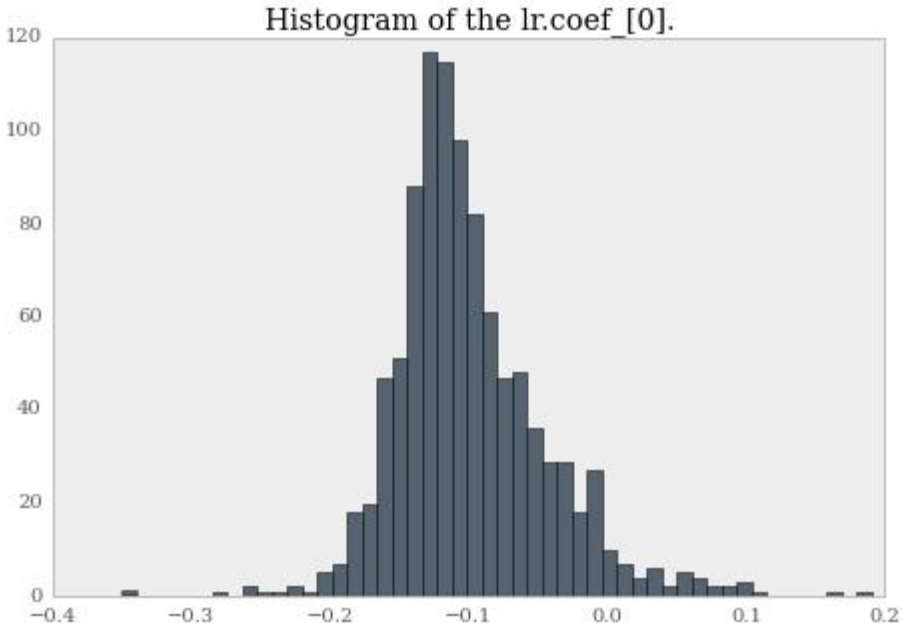

GBR Residuals vs LR Residuals
GBR Residuals
LR Residuals
120
100
80
60
40
20
0
−40
−30
−20
−10
0
10
20
30
40

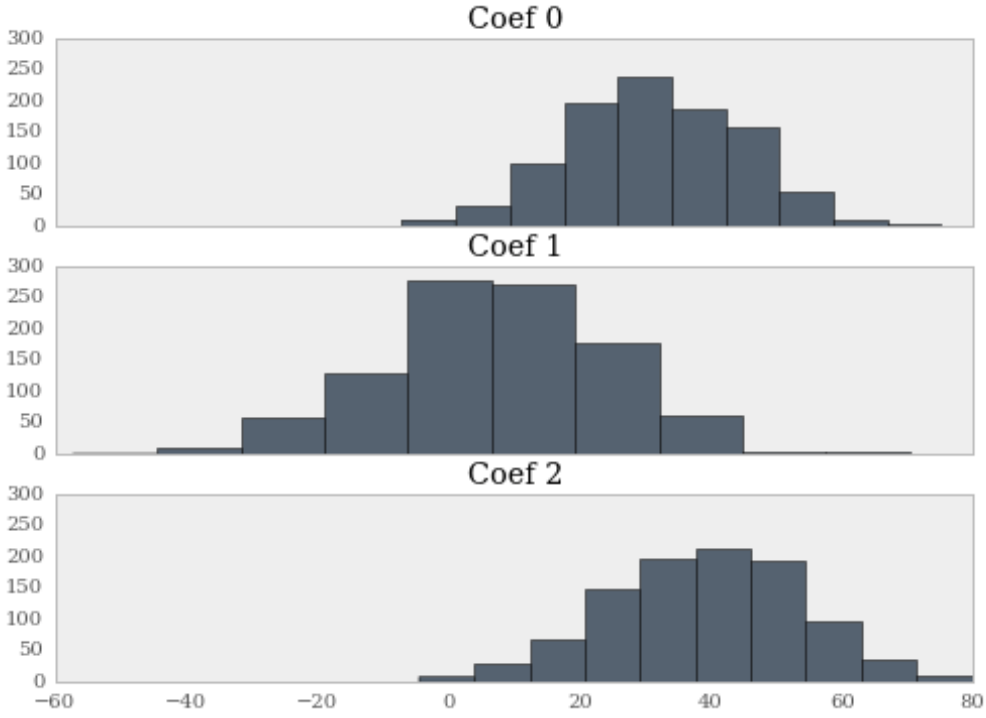

Residuals at Various Numbers of Estimators
n_estimators: 800
n_estimators: 450
n_estimators: 100
160
140
120
100
80
60
40
20
0
−30
−20
−10
0
10
20
30
40

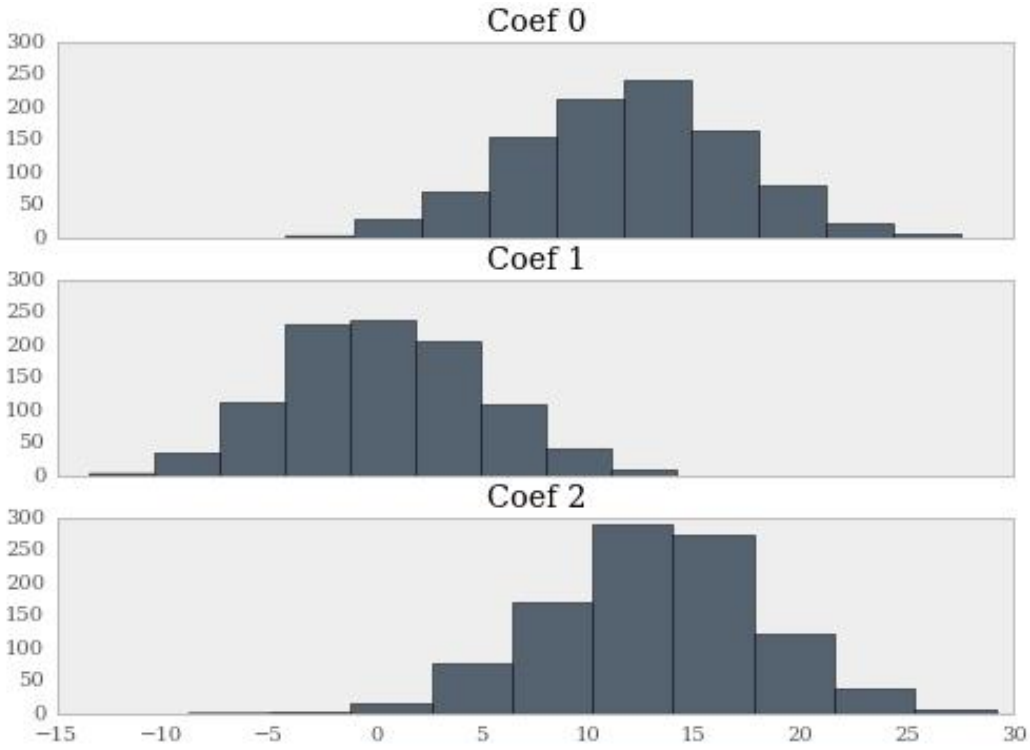

Histogram of Residuals.
70
60
50
40
30
20
10
0
−20
−10
0
10
20
30

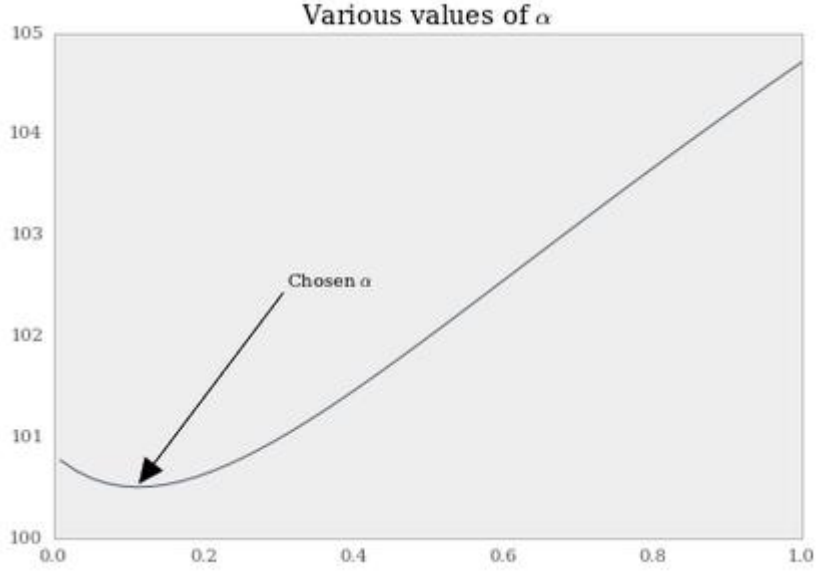

Histogram of Residuals.
70
60
50
40
30
20
10
0
−20
−10
0
10
20
30

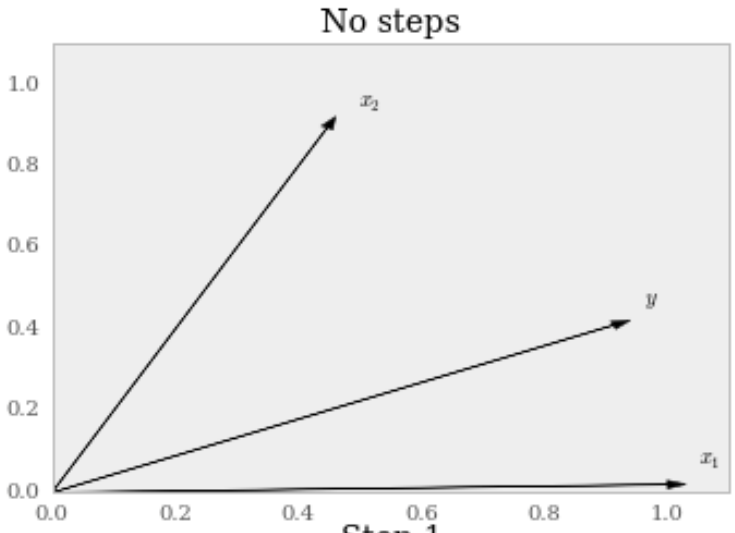

Probability Plot
Ordered Values
Quantiles
30
20
10
0
−10
−20
−3
−2
−1
0
1
2
3
$R^2$ =0.9476

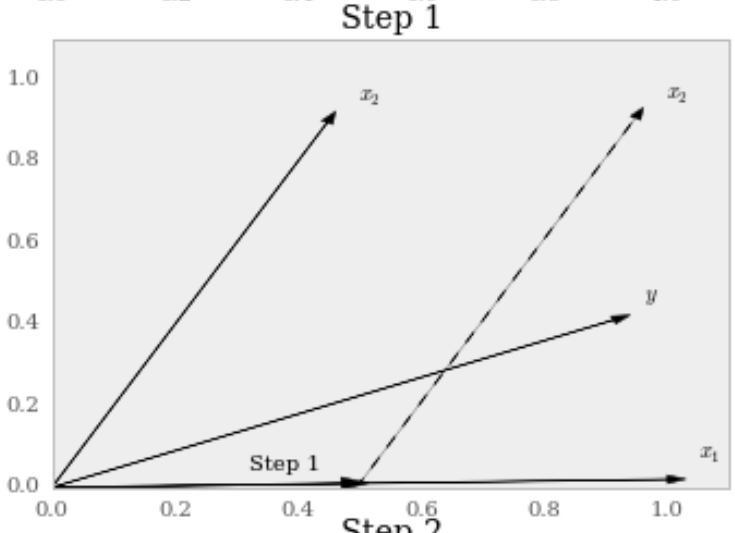

Histogram of the lr.coef_[0].
120
100
80
60
40
20
0
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2

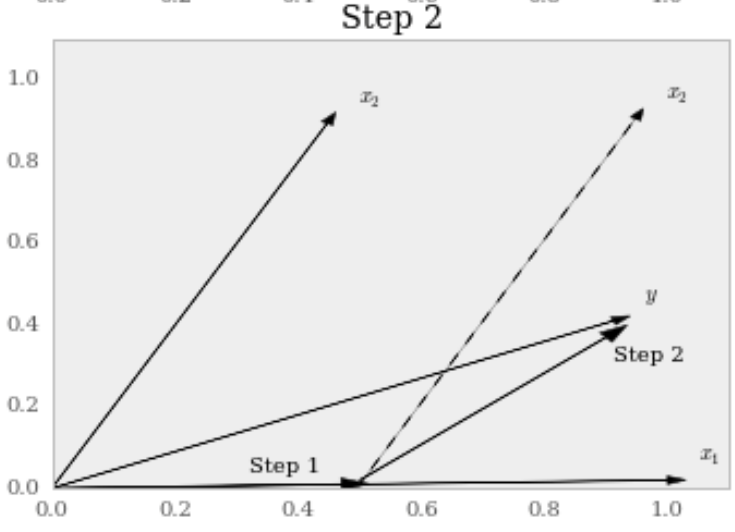

Coef 0
Coef 1
Coef 2
300
250
200
150
100
50
0
−60
−40
−20
0
20
40
60
80

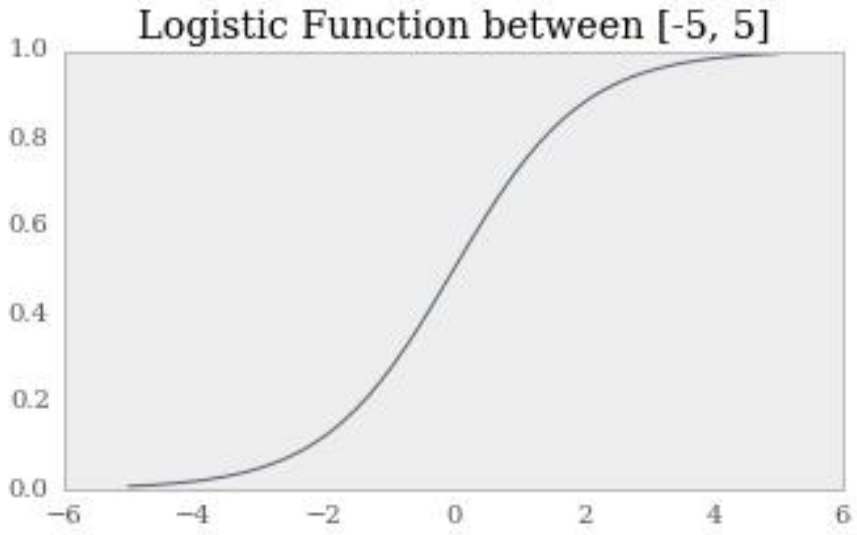

Coef 0
Coef 1
Coef 2
300
250
200
150
100
50
0
−15
−10
−5
0
5
10
15
20
25
30

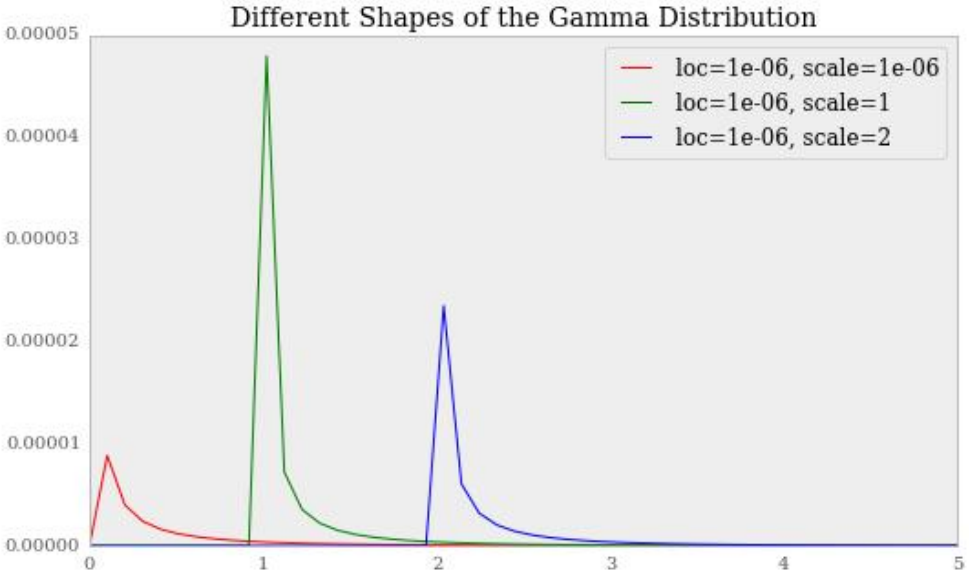

Various values of α
Chosen α
105
104
103
102
101
100
0.0
0.2
0.4
0.6
0.8
1.0

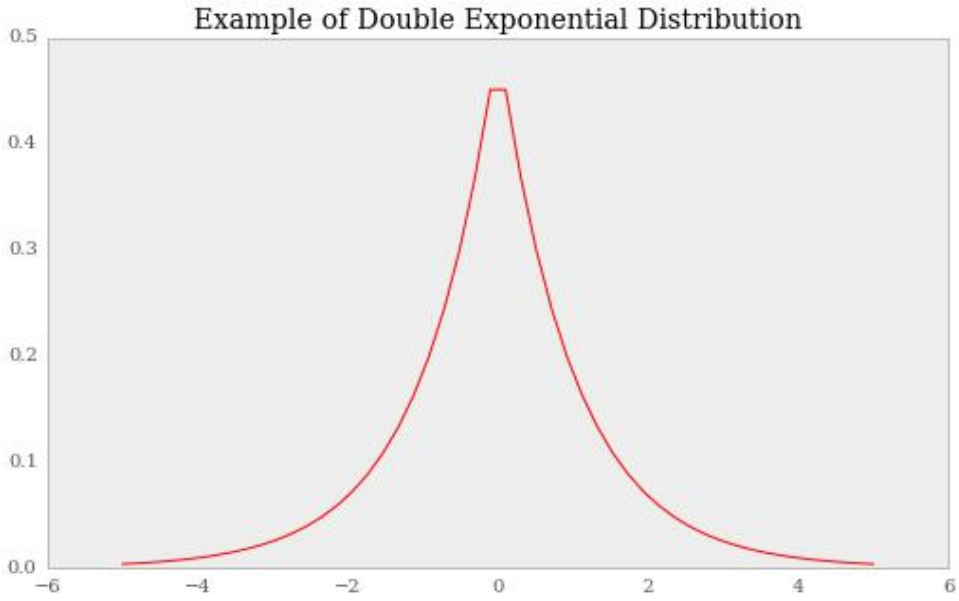

No steps
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
$x_2$
$y$
$x_1$

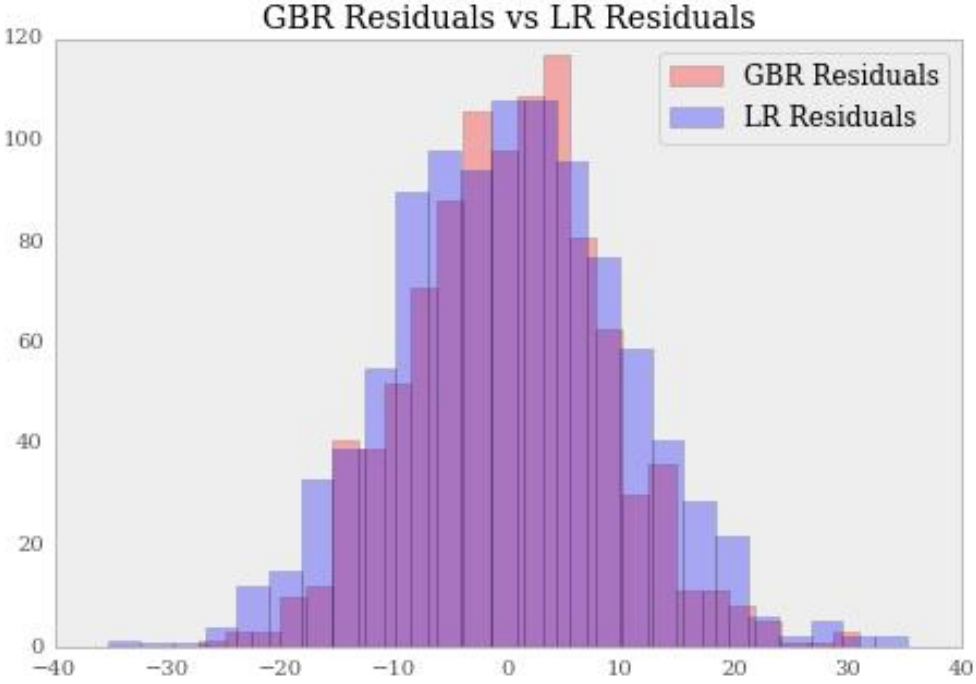

Step 1
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
$x_2$
$x_2$
$y$
Step 1
$x_1$

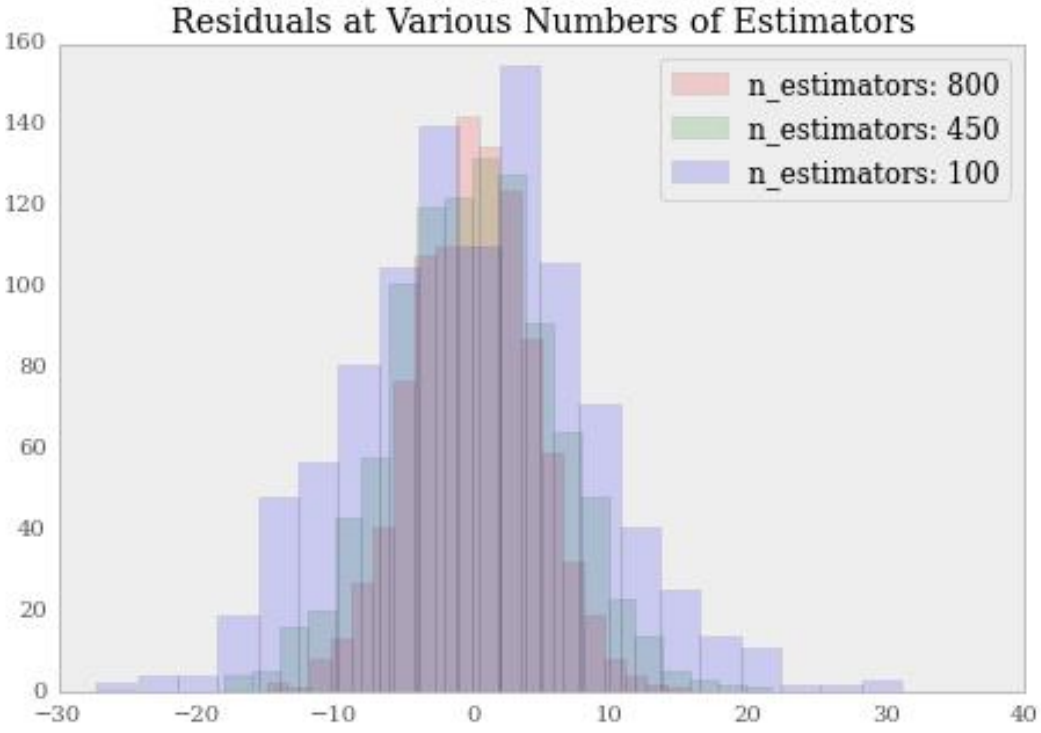

Step 2
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
$x_2$
$x_2$
$y$
Step 2
Step 1
$x_1$


Logistic Function between [-5, 5]
1.0
0.8
0.6
0.4
0.2
0.0
−6
−4
−2
0
2
4
6

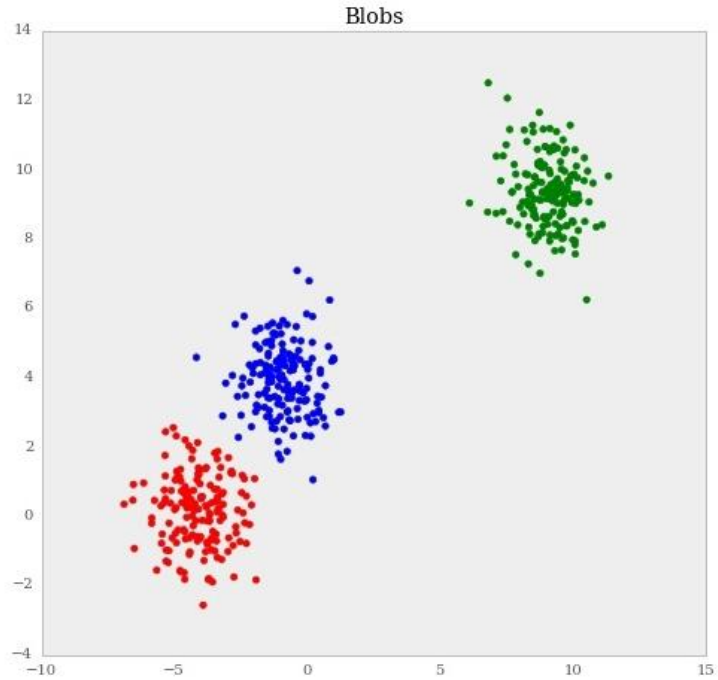

Different Shapes of the Gamma Distribution
loc=1e-06, scale=1e-06
loc=1e-06, scale=1
loc=1e-06, scale=2
0.00005
0.00004
0.00003
0.00002
0.00001
0.00000
0
1
2
3
4
5

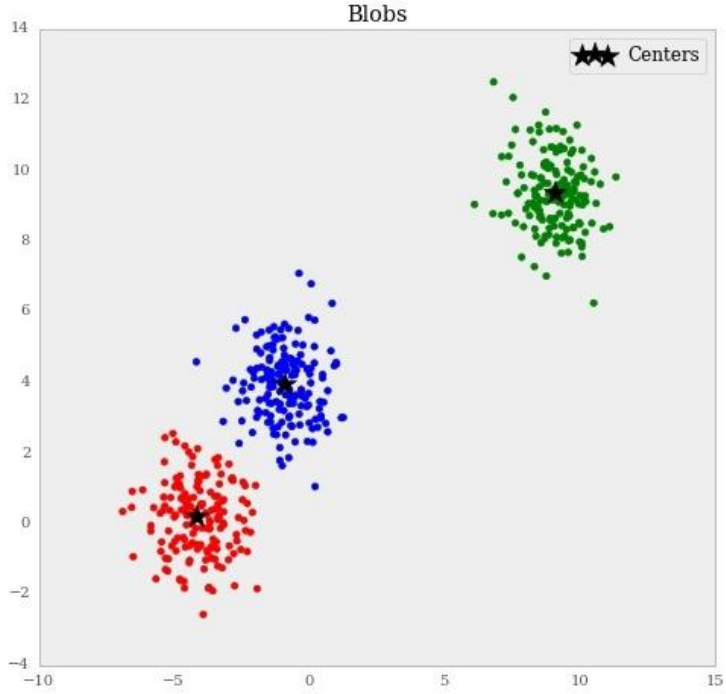

Example of Double Exponential Distribution
0.5
0.4
0.3
0.2
0.1
0.0
−6
−4
−2
0
2
4
6

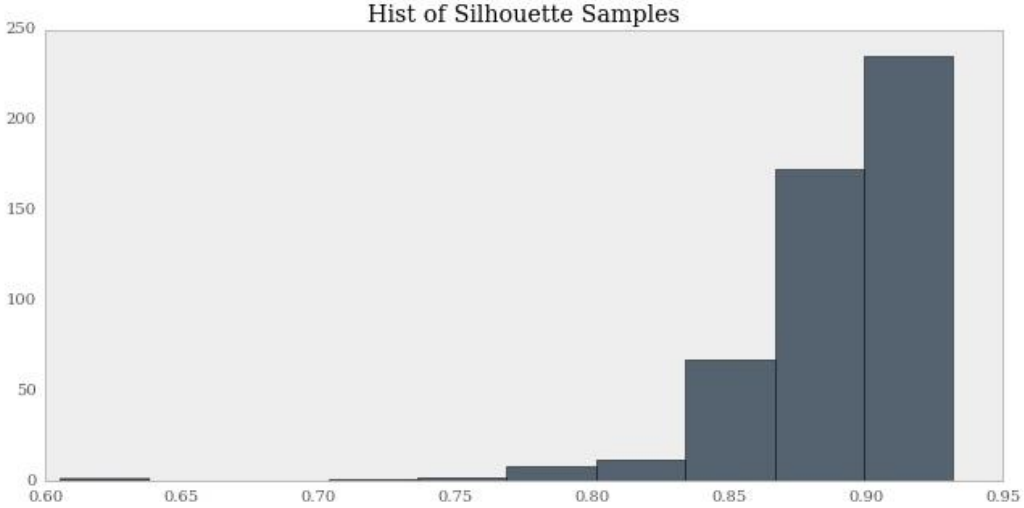

GBR Residuals vs LR Residuals
GBR Residuals
LR Residuals
120
100
80
60
40
20
0
−40
−30
−20
−10
0
10
20
30
40

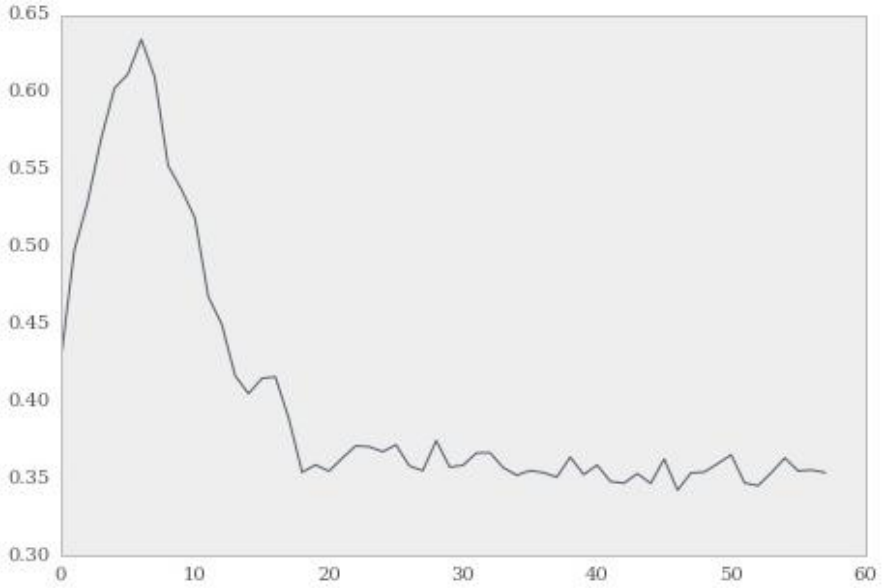

Residuals at Various Numbers of Estimators
n_estimators: 800
n_estimators: 450
n_estimators: 100
160
140
120
100
80
60
40
20
0
−30
−20
−10
0
10
20
30
40

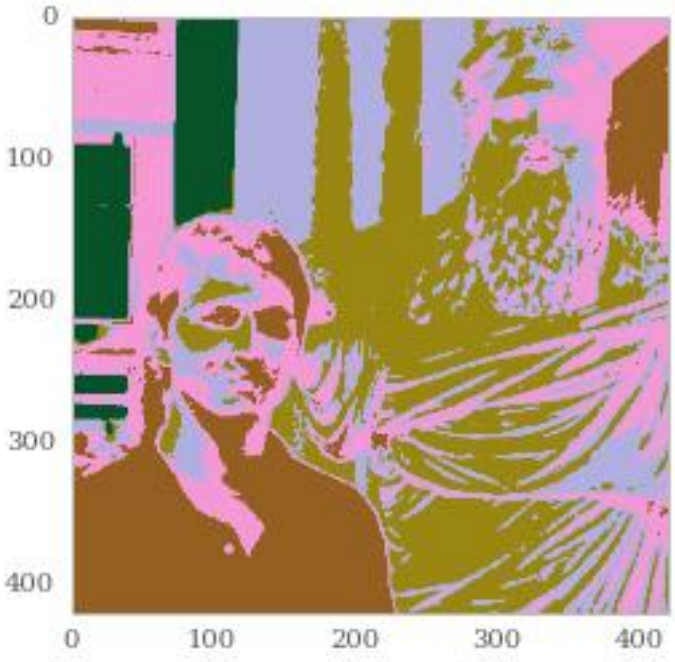
# Chapter 3

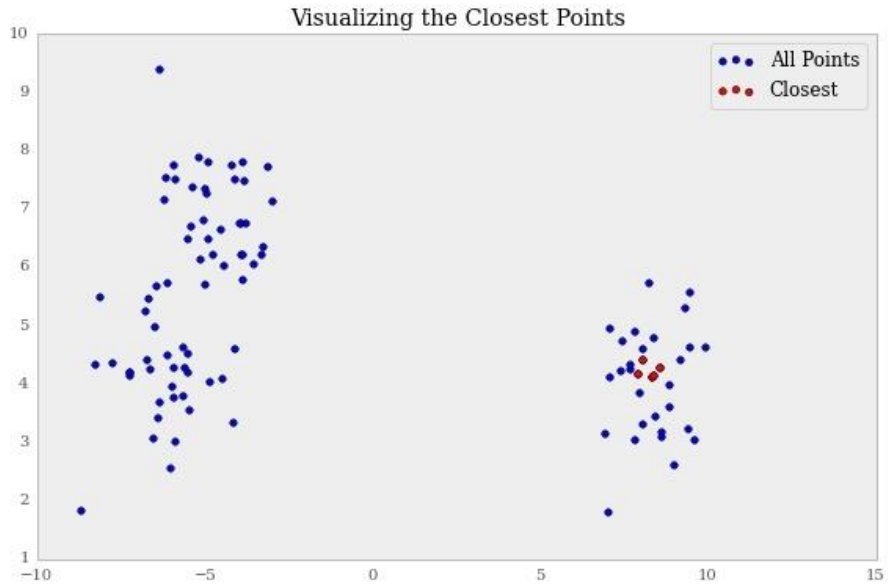

Blobs
14
12
10
8
6
4
2
0
−2
−4
−10
−5
0
5
10
15

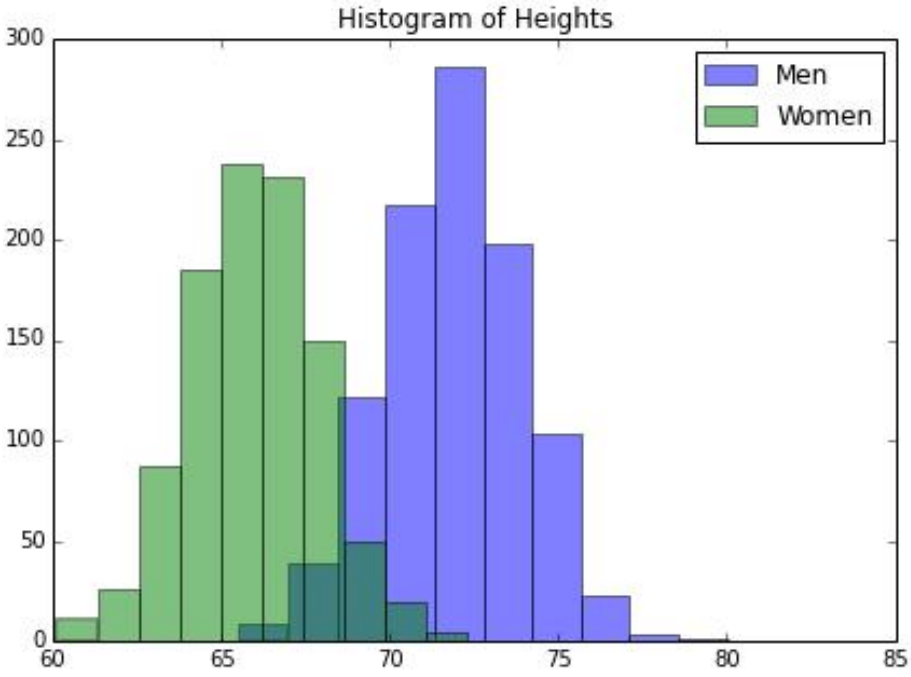

Blobs
Centers
14
12
10
8
6
4
2
0
−2
−4
−10
−5
0
5
10
15

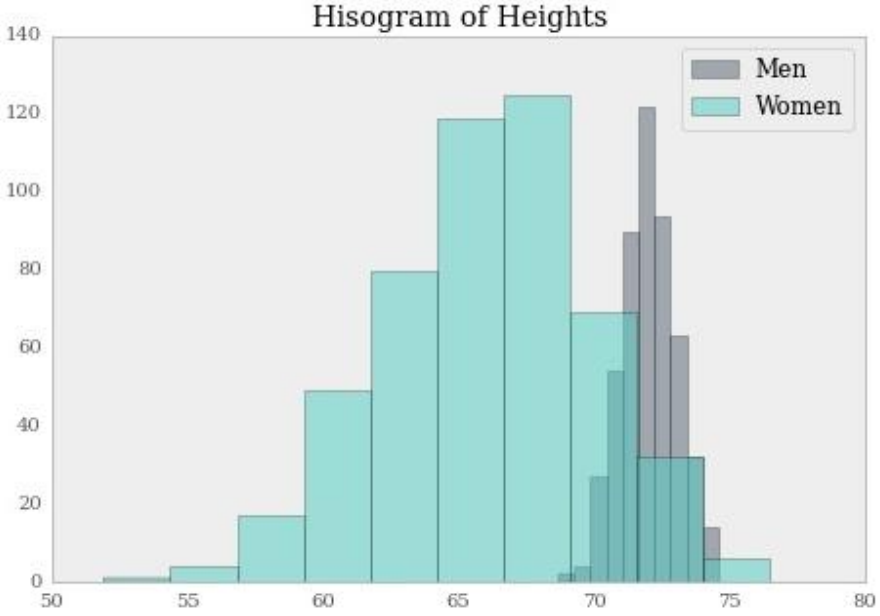

Hist of Silhouette Samples
250
200
150
100
50
0
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.95

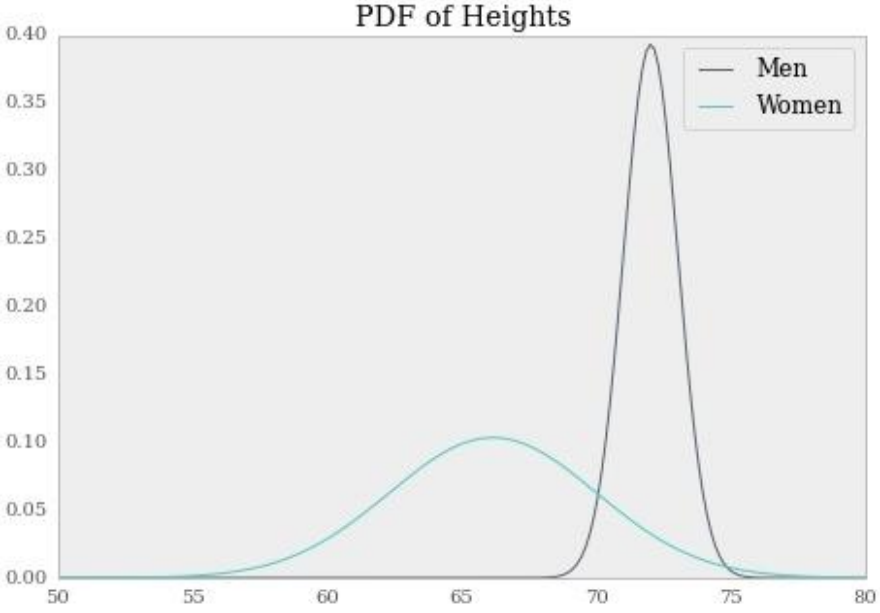

0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0
10
20
30
40
50
60

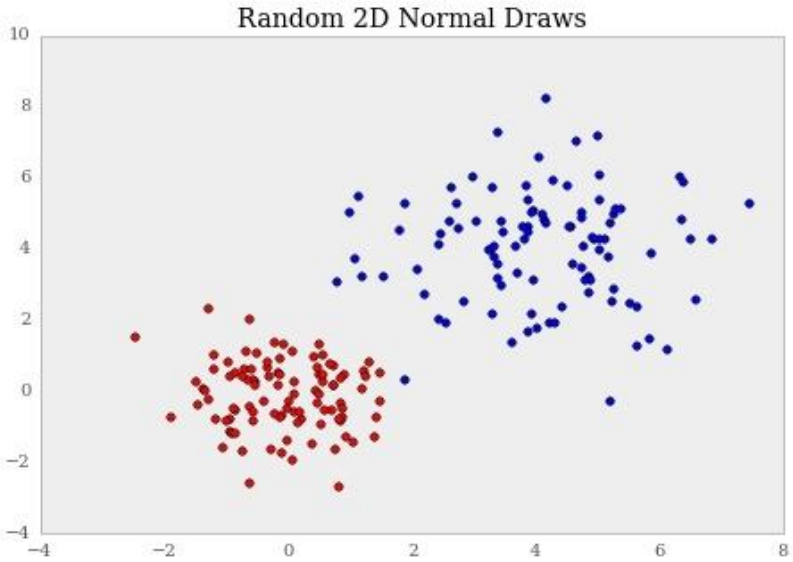

0
100
200
300
400
0
100
200
300
400

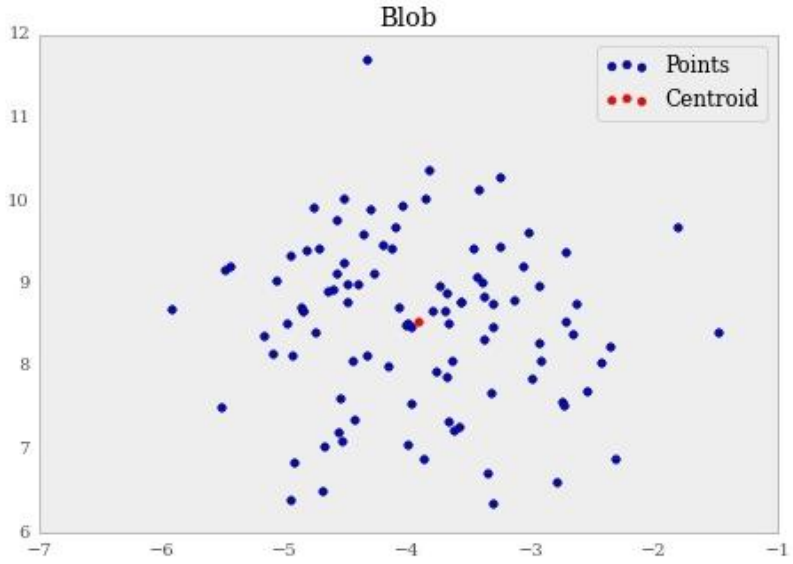

Visualizing the Closest Points
All Points
Closest
10
9
8
7
6
5
4
3
2
1
−10
−5
0
5
10
15

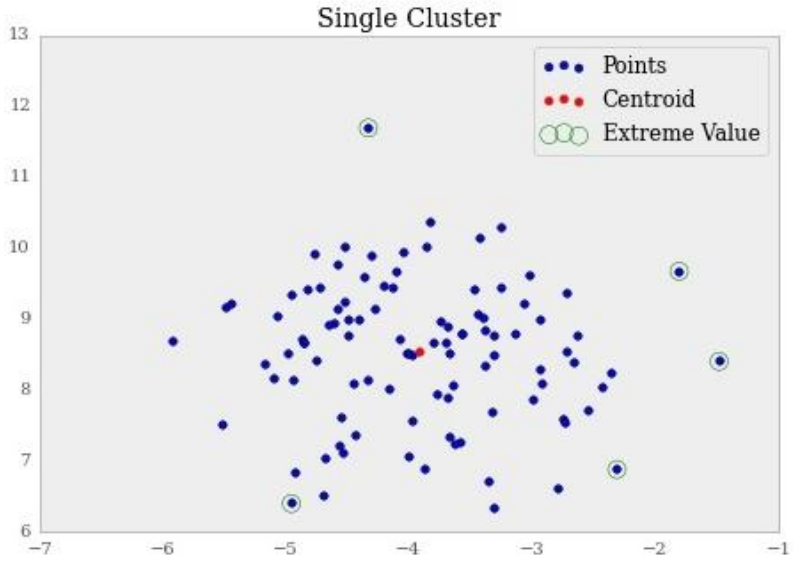

Histogram of Heights
Men
Women
300
250
200
150
100
50
0
60
65
70
75
80
85

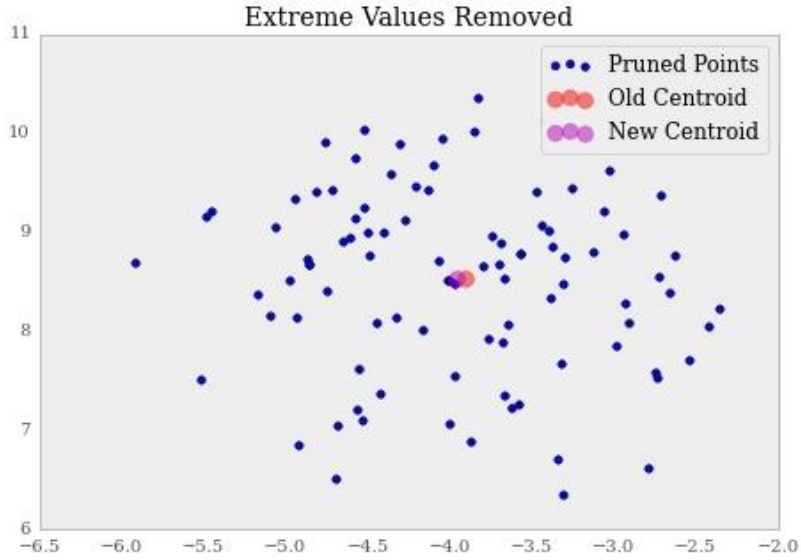

Hisogram of Heights
Men
Women
140
120
100
80
60
40
20
0
50
55
60
65
70
75
80

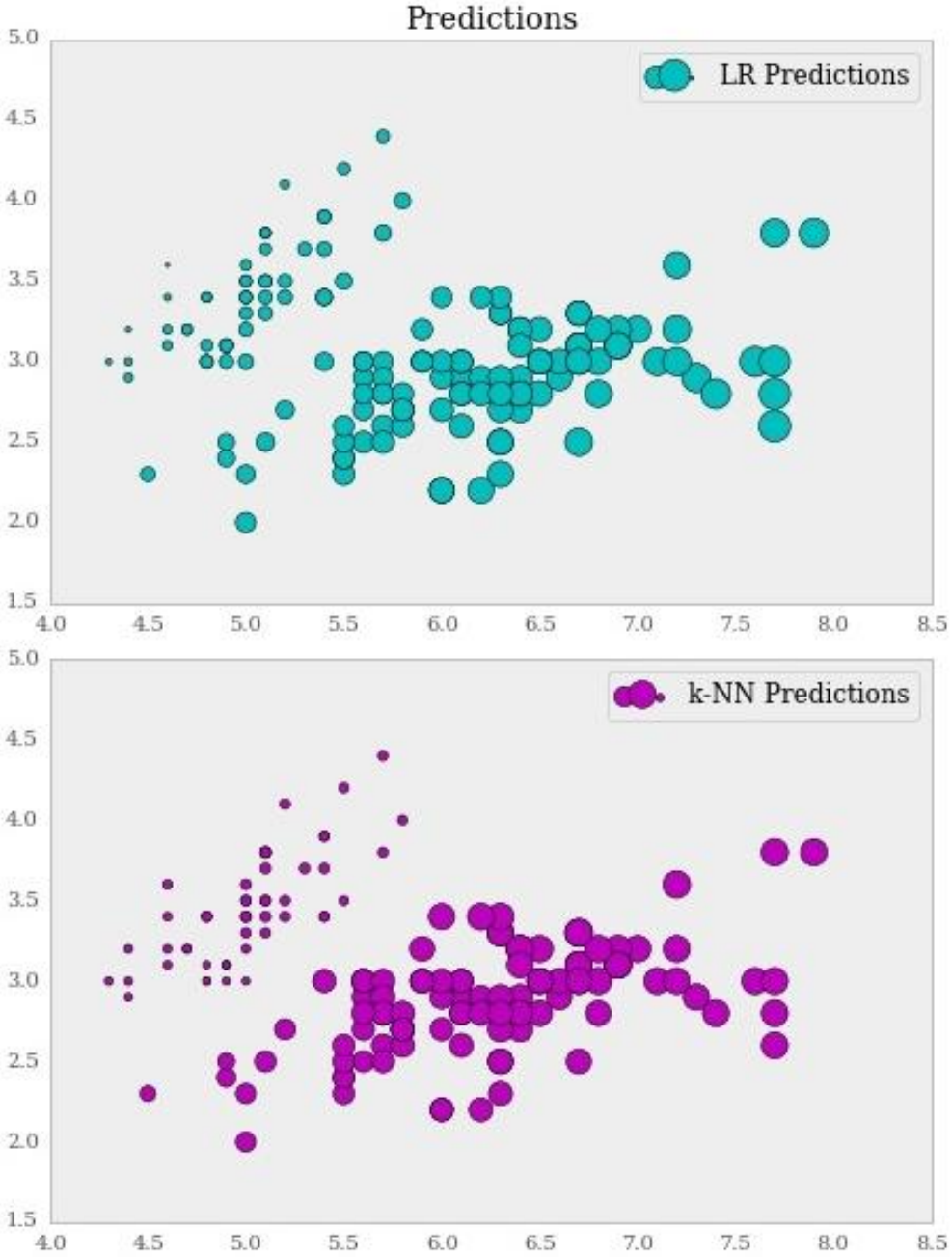

PDF of Heights
Men
Women
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
50
55
60
65
70
75
80

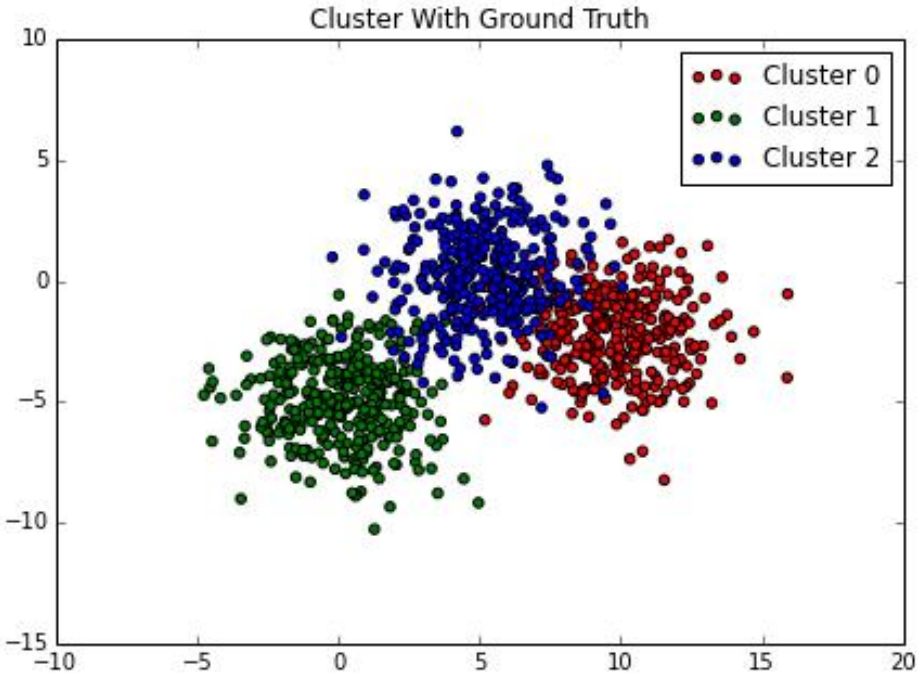

Random 2D Normal Draws
10
8
6
4
2
0
−2
−4
−4
−2
0
2
4
6
8

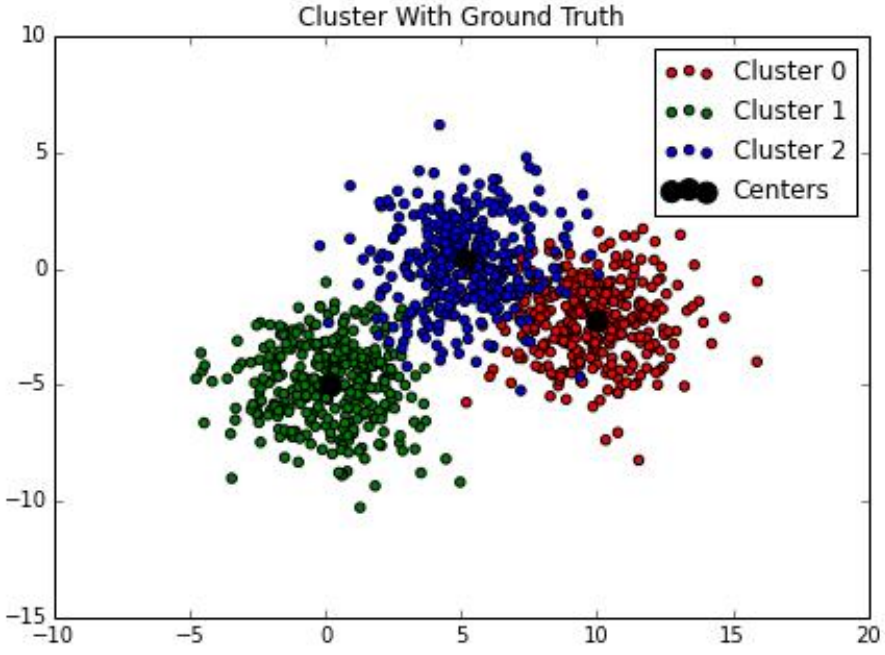

Blob
Points
Centroid
12
11
10
9
8
7
6
−7
−6
−5
−4
−3
−2
−1


Single Cluster
Points
Centroid
Extreme Value
13
12
11
10
9
8
7
6
−7
−6
−5
−4
−3
−2
−1

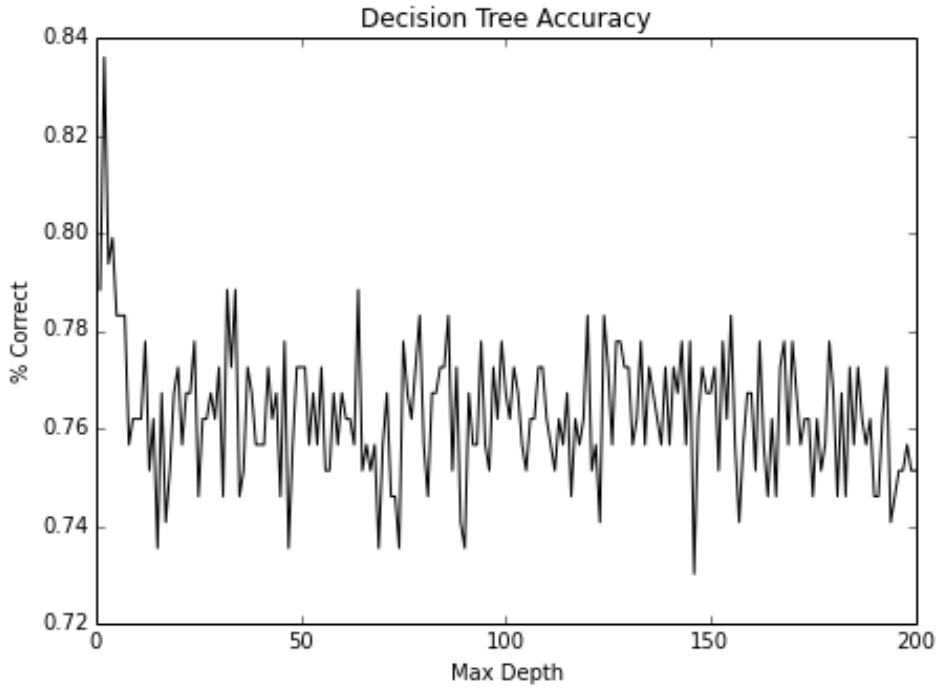

Extreme Values Removed
Pruned Points
Old Centroid
New Centroid
11
10
9
8
7
6
−6.5
−6.0
−5.5
−5.0
−4.5
−4.0
−3.5
−3.0
−2.5
−2.0

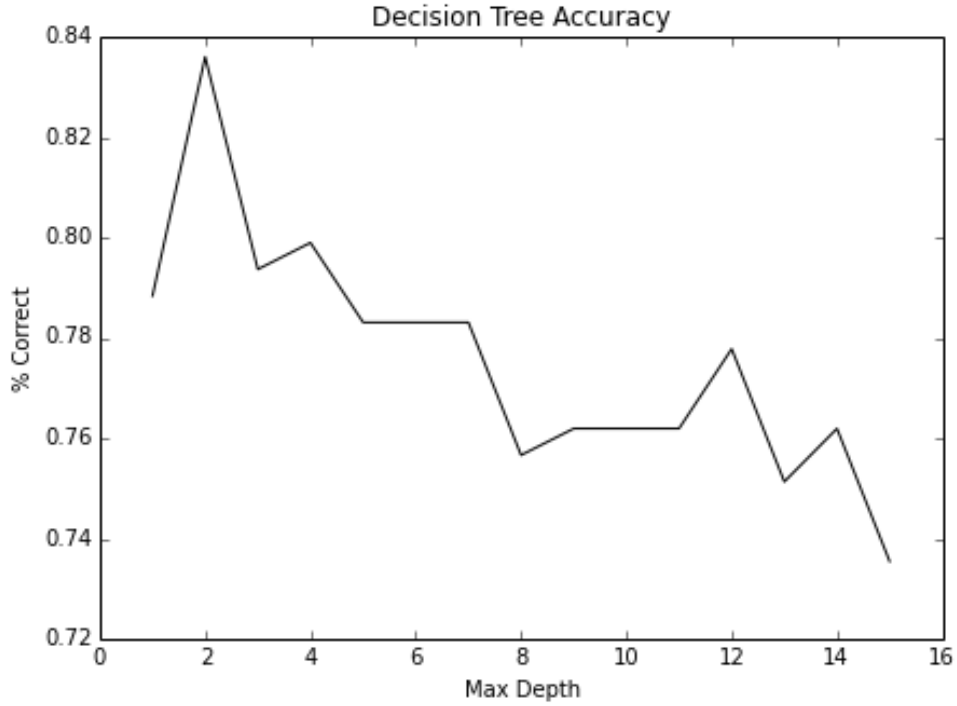

Predictions
LR Predictions
k-NN Predictions
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
8.5

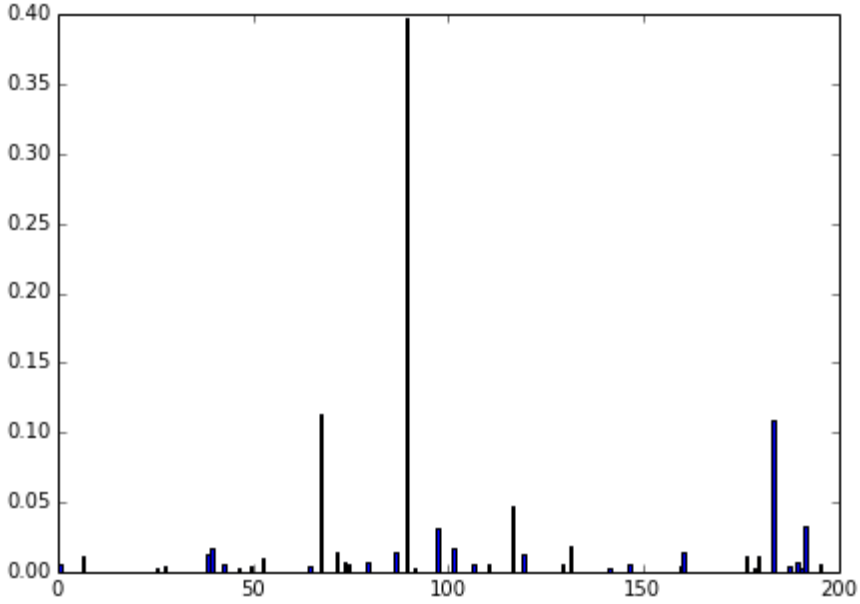

Cluster With Ground Truth
Cluster 0
Cluster 1
Cluster 2
10
5
0
−5
−10
−15
−10
−5
0
5
10
15
20

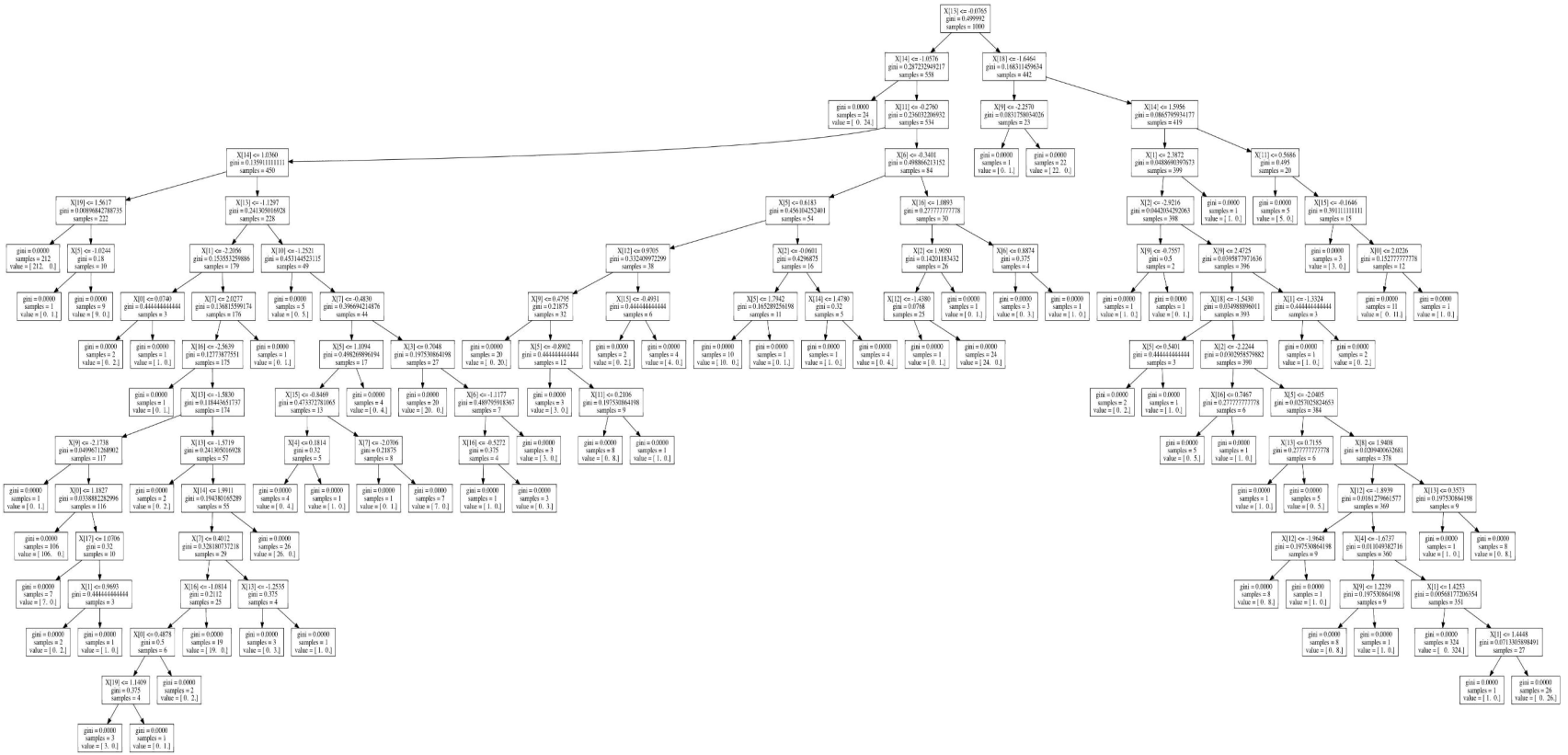

Cluster With Ground Truth
Cluster 0
Cluster 1
Cluster 2
Centers
10
5
0
−5
−10
−15
−10
−5
0
5
10
15
20

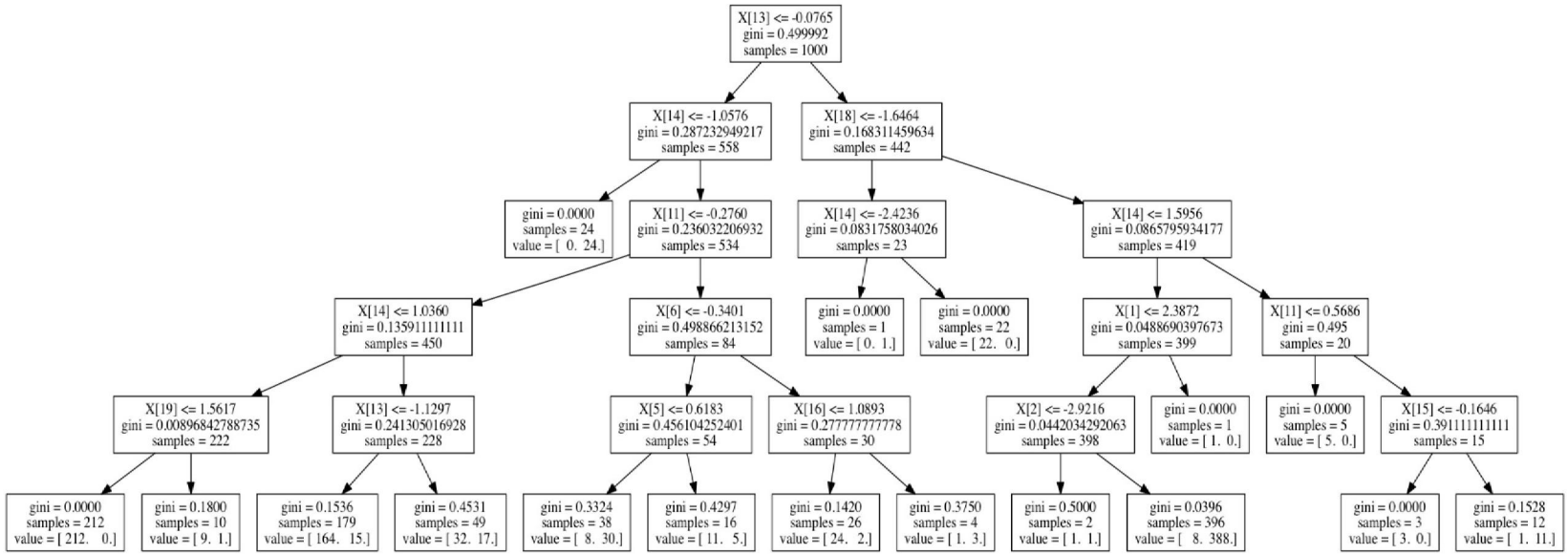
# Chapter 4

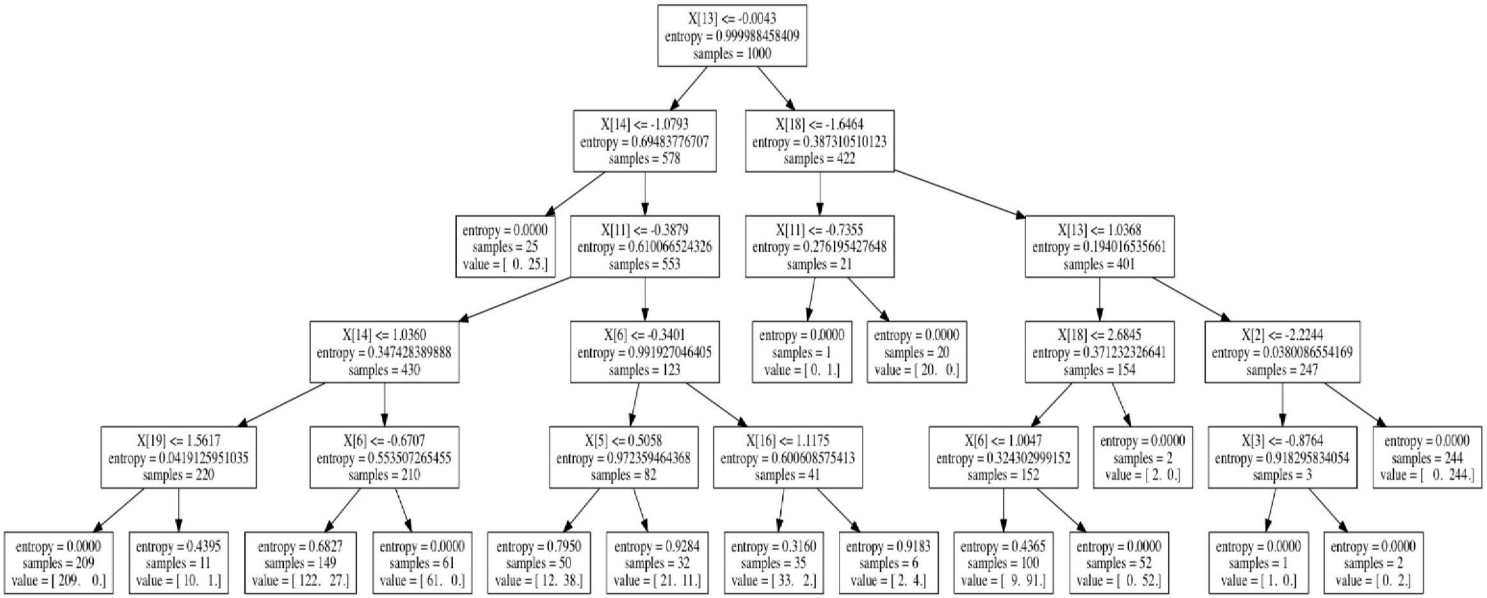

Decision Tree Accuracy
% Correct
Max Depth
0.84
0.82
0.80
0.78
0.76
0.74
0.72
0
50
100
150
200

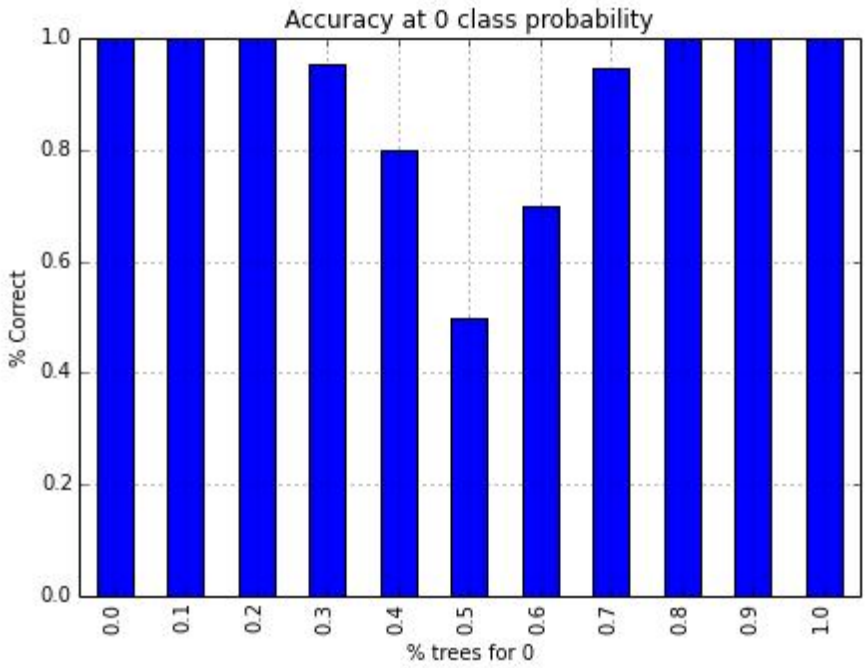

Decision Tree Accuracy
% Correct
Max Depth
0.84
0.82
0.80
0.78
0.76
0.74
0.72
0
2
4
6
8
10
12
14
16

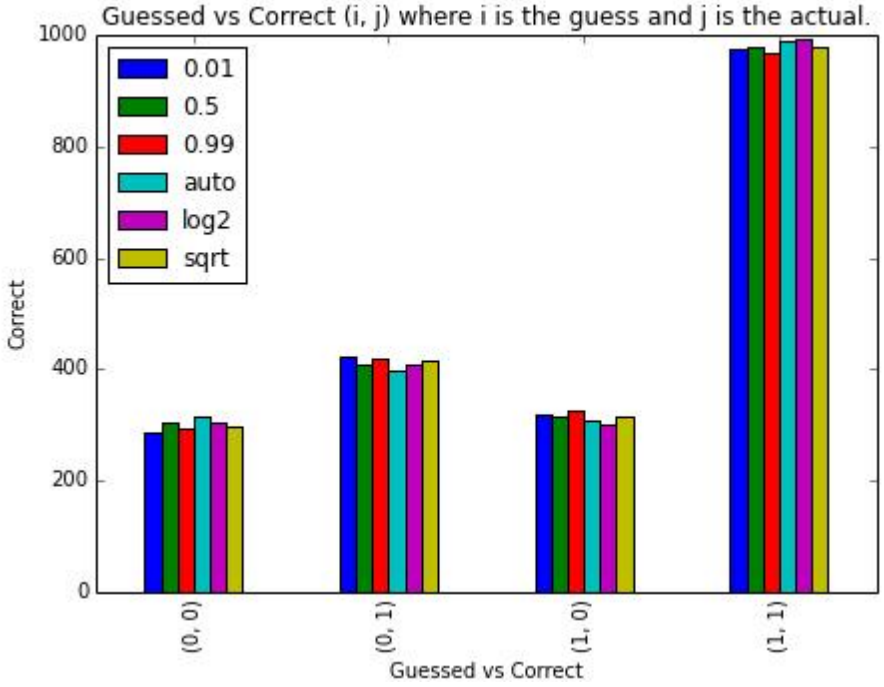

0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
50
100
150
200

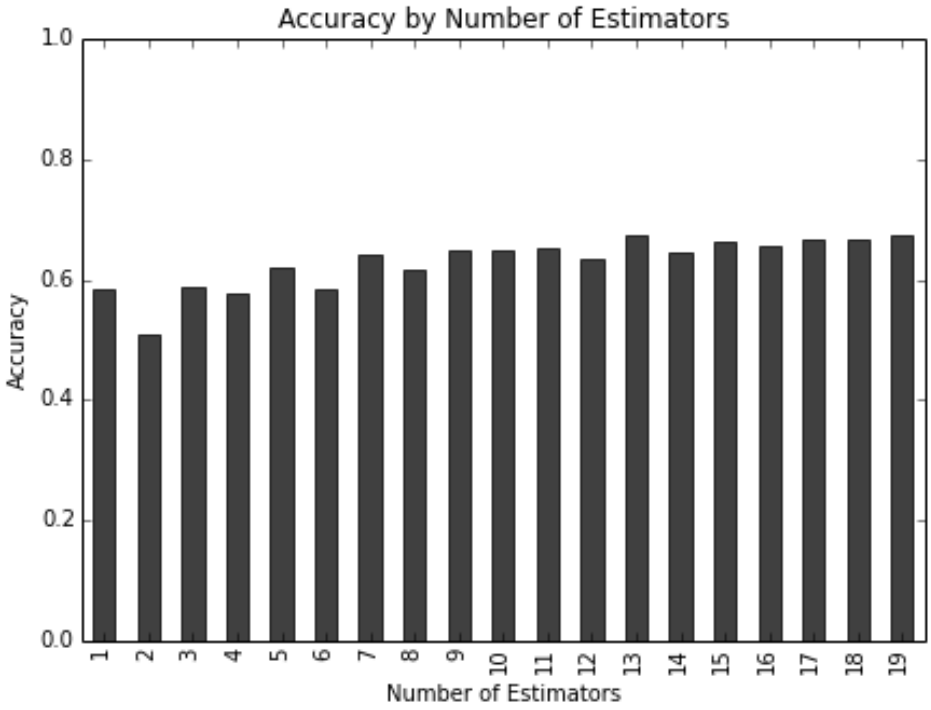

X[13] <= -0.0765
gini = 0.499992
samples = 1000
X[14] <= -1.0576
gini = 0.287232949217
samples = 558
X[18] <= -1.6464
gini = 0.183314159634
samples = 442
gini = 0.0000
samples = 24
value = [ 0. 24.]
X[11] <= -0.2760
gini = 0.236032206932
samples = 534
X[9] <= -2.2570
gini = 0.0831758034026
samples = 25
X[14] <= 1.5956
gini = 0.0865795934177
samples = 419
gini = 0.0000
samples = 1
value = [ 0. 1.]
gini = 0.0000
samples = 22
value = [ 22. 0.]
X[14] <= 1.0360
gini = 0.135911111111
samples = 450
X[6] <= -0.3401
gini = 0.498866213152
samples = 84
X[1] <= 2.3872
gini = 0.0488603597673
samples = 399
X[11] <= 0.5686
gini = 0.495
samples = 20
X[19] <= 1.5617
gini = 0.00699842780735
samples = 222
X[13] <= -1.1297
gini = 0.241309016928
samples = 228
X[5] <= 0.6183
gini = 0.458104252401
samples = 54
X[16] <= 1.0893
gini = 0.277777777778
samples = 30
X[2] <= -2.9216
gini = 0.0442054270263
samples = 398
gini = 0.0000
samples = 1
value = [ 1. 0.]
gini = 0.0000
samples = 5
value = [ 5. 0.]
X[15] <= -0.1646
gini = 0.391111111111
samples = 15
gini = 0.0000
samples = 212
value = [ 212. 0.]
X[5] <= -1.0244
gini = 0.18
samples = 10
X[11] <= -2.2056
gini = 0.353532259886
samples = 179
X[10] <= -1.2521
gini = 0.45314452115
samples = 49
X[12] <= 0.9305
gini = 0.332409972299
samples = 38
X[2] <= -0.1601
gini = 0.4296875
samples = 16
X[2] <= 1.9050
gini = 0.142011834332
samples = 26
X[6] <= 0.8874
gini = 0.375
samples = 4
X[9] <= -0.7557
gini = 0.5
samples = 2
X[9] <= 2.4725
gini = 0.0395817797636
samples = 396
gini = 0.0000
samples = 3
value = [ 3. 0.]
X[10] <= 2.0226
gini = 0.152777777778
samples = 12
gini = 0.0000
samples = 1
value = [ 1. 0.]
gini = 0.0000
samples = 9
value = [ 9. 0.]
X[0] <= 0.0740
gini = 0.444444444444
samples = 3
X[7] <= 2.0277
gini = 0.136815599174
samples = 176
gini = 0.0000
samples = 5
value = [ 0. 5.]
X[7] <= -0.4830
gini = 0.396694214876
samples = 44
X[9] <= 0.4795
gini = 0.21875
samples = 32
X[5] <= -0.4931
gini = 0.444444444444
samples = 6
X[5] <= 1.7942
gini = 0.165289256198
samples = 11
X[14] <= 1.4780
gini = 0.32
samples = 5
X[12] <= -1.4380
gini = 0.0768
samples = 25
gini = 0.0000
samples = 1
value = [ 0. 1.]
gini = 0.0000
samples = 3
value = [ 0. 3.]
gini = 0.0000
samples = 1
value = [ 1. 0.]
gini = 0.0000
samples = 1
value = [ 1. 0.]
gini = 0.0000
samples = 1
value = [ 1. 0.]
X[18] <= -1.5430
gini = 0.0349888896011
samples = 393
X[1] <= -1.3324
gini = 0.444444444444
samples = 3
gini = 0.0000
samples = 11
value = [ 0. 11.]
gini = 0.0000
samples = 1
value = [ 1. 0.]
gini = 0.0000
samples = 2
value = [ 0. 2.]
gini = 0.0000
samples = 1
value = [ 1. 0.]
X[16] <= -2.5639
gini = 0.127738775511
samples = 175
X[5] <= 1.1094
gini = 0.498269896194
samples = 17
X[3] <= 0.7048
gini = 0.197530864198
samples = 27
X[5] <= -0.8902
gini = 0.444444444444
samples = 12
gini = 0.0000
samples = 6
value = [ 6. 0.]
gini = 0.0000
samples = 4
value = [ 4. 0.]
gini = 0.0000
samples = 1
value = [ 1. 0.]
gini = 0.0000
samples = 1
value = [ 1. 0.]
gini = 0.0000
samples = 4
value = [ 0. 4.]
gini = 0.0000
samples = 24
value = [ 24. 0.]
X[5] <= 0.5401
gini = 0.444444444444
samples = 3
X[2] <= -2.2244
gini = 0.0192595579882
samples = 390
gini = 0.0000
samples = 1
value = [ 1. 0.]
gini = 0.0000
samples = 2
value = [ 0. 2.]
X[13] <= -1.5830
gini = 0.118445631737
samples = 174
X[15] <= -0.8469
gini = 0.473372781065
samples = 13
gini = 0.0000
samples = 4
value = [ 0. 4.]
gini = 0.0000
samples = 20
value = [ 20. 0.]
X[8] <= -1.1177
gini = 0.489795918367
samples = 7
gini = 0.0000
samples = 3
value = [ 3. 0.]
X[11] <= 0.2106
gini = 0.197530864198
samples = 9
gini = 0.0000
samples = 1
value = [ 1. 0.]
X[16] <= 0.7467
gini = 0.277777777778
samples = 6
X[5] <= -2.0405
gini = 0.0370258424653
samples = 384
gini = 0.0000
samples = 5
value = [ 0. 5.]
gini = 0.0000
samples = 1
value = [ 1. 0.]
X[9] <= 0.2738
gini = 0.0499671268802
samples = 117
X[13] <= -1.5719
gini = 0.241305016928
samples = 57
X[4] <= 0.1814
gini = 0.32
samples = 5
X[7] <= -2.0706
gini = 0.21875
samples = 8
X[16] <= -0.5272
gini = 0.375
samples = 4
gini = 0.0000
samples = 3
value = [ 3. 0.]
gini = 0.0000
samples = 8
value = [ 0. 8.]
gini = 0.0000
samples = 1
value = [ 1. 0.]
X[13] <= 0.7155
gini = 0.277777777778
samples = 6
X[8] <= 1.9408
gini = 0.0209400632681
samples = 378
gini = 0.0000
samples = 7
value = [ 7. 0.]
X[0] <= 1.1827
gini = 0.0338882282996
samples = 110
gini = 0.0000
samples = 2
value = [ 0. 2.]
X[14] <= 1.4911
gini = 0.19438016529
samples = 55
gini = 0.0000
samples = 4
value = [ 0. 4.]
gini = 0.0000
samples = 1
value = [ 1. 0.]
gini = 0.0000
samples = 1
value = [ 0. 1.]
gini = 0.0000
samples = 7
value = [ 7. 0.]
gini = 0.0000
samples = 1
value = [ 1. 0.]
gini = 0.0000
samples = 3
value = [ 0. 3.]
gini = 0.0000
samples = 1
value = [ 1. 0.]
gini = 0.0000
samples = 5
value = [ 0. 5.]
X[12] <= -1.8939
gini = 0.0181279661577
samples = 369
X[13] <= 0.3573
gini = 0.197530864198
samples = 9
gini = 0.0000
samples = 106
value = [ 106. 0.]
X[17] <= 1.0706
gini = 0.32
samples = 10
X[7] <= 0.4012
gini = 0.328180732218
samples = 29
gini = 0.0000
samples = 26
value = [ 26. 0.]
X[12] <= -1.9648
gini = 0.197530864198
samples = 9
X[4] <= -1.6737
gini = 0.0110491852716
samples = 360
gini = 0.0000
samples = 1
value = [ 1. 0.]
gini = 0.0000
samples = 8
value = [ 0. 8.]
gini = 0.0000
samples = 7
value = [ 7. 0.]
X[1] <= 0.9693
gini = 0.444444444444
samples = 3
X[16] <= -1.0814
gini = 0.2112
samples = 25
X[13] <= -1.2555
gini = 0.375
samples = 4
gini = 0.0000
samples = 8
value = [ 0. 8.]
gini = 0.0000
samples = 1
value = [ 1. 0.]
X[9] <= 1.2239
gini = 0.197530864198
samples = 9
X[1] <= 1.4253
gini = 0.00568177286354
samples = 351
gini = 0.0000
samples = 2
value = [ 0. 2.]
gini = 0.0000
samples = 1
value = [ 1. 0.]
X[0] <= 0.4878
gini = 0.5
samples = 6
gini = 0.0000
samples = 19
value = [ 19. 0.]
gini = 0.0000
samples = 3
value = [ 0. 3.]
gini = 0.0000
samples = 1
value = [ 1. 0.]
gini = 0.0000
samples = 8
value = [ 0. 8.]
gini = 0.0000
samples = 1
value = [ 1. 0.]
gini = 0.0000
samples = 324
value = [ 0. 324.]
X[1] <= 1.4448
gini = 0.0771330589849
samples = 27
X[9] <= 1.1409
gini = 0.375
samples = 4
gini = 0.0000
samples = 2
value = [ 0. 2.]
gini = 0.0000
samples = 1
value = [ 1. 0.]
gini = 0.0000
samples = 26
value = [ 0. 26.]
gini = 0.0000
samples = 3
value = [ 3. 0.]
gini = 0.0000
samples = 1
value = [ 0. 1.]

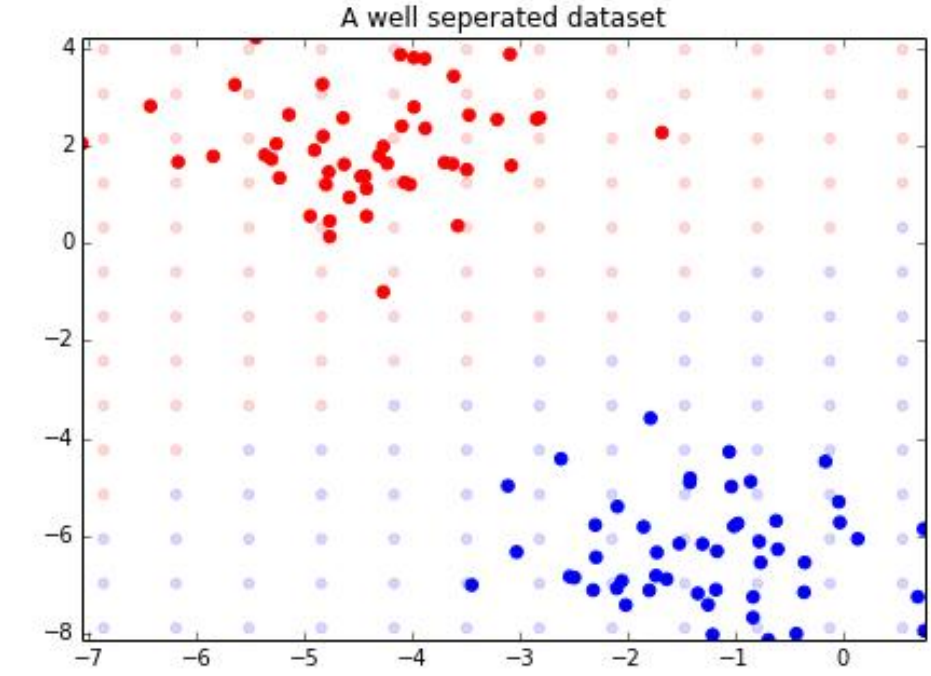

X[13] <= -0.0765
gini = 0.499992
samples = 1000
X[14] <= -1.0576
gini = 0.287232949217
samples = 558
X[18] <= -1.6464
gini = 0.168311459634
samples = 442
gini = 0.0000
samples = 24
value = [ 0. 24.]
X[11] <= -0.2760
gini = 0.236032206932
samples = 534
X[14] <= -2.4236
gini = 0.0831758034026
samples = 23
X[14] <= 1.5956
gini = 0.0865795934177
samples = 419
X[14] <= 1.0360
gini = 0.135911111111
samples = 450
X[6] <= -0.3401
gini = 0.498866213152
samples = 84
gini = 0.0000
samples = 1
value = [ 0. 1.]
gini = 0.0000
samples = 22
value = [ 22. 0.]
X[1] <= 2.3872
gini = 0.0488690397673
samples = 399
X[11] <= 0.5686
gini = 0.495
samples = 20
X[19] <= 1.5617
gini = 0.00896842788735
samples = 222
X[13] <= -1.1297
gini = 0.241305016928
samples = 228
X[5] <= 0.6183
gini = 0.456104252401
samples = 54
X[16] <= 1.0893
gini = 0.277777777778
samples = 30
X[2] <= -2.9216
gini = 0.0442034292063
samples = 398
gini = 0.0000
samples = 1
value = [ 1. 0.]
gini = 0.0000
samples = 5
value = [ 5. 0.]
X[15] <= -0.1646
gini = 0.391111111111
samples = 15
gini = 0.0000
samples = 212
value = [ 212. 0.]
gini = 0.1800
samples = 10
value = [ 9. 1.]
gini = 0.1536
samples = 179
value = [ 164. 15.]
gini = 0.4531
samples = 49
value = [ 32. 17.]
gini = 0.3324
samples = 38
value = [ 8. 30.]
gini = 0.4297
samples = 16
value = [ 11. 5.]
gini = 0.1420
samples = 26
value = [ 24. 2.]
gini = 0.3750
samples = 4
value = [ 1. 3.]
gini = 0.5000
samples = 2
value = [ 1. 1.]
gini = 0.0396
samples = 396
value = [ 8. 388.]
gini = 0.0000
samples = 3
value = [ 3. 0.]
gini = 0.1528
samples = 12
value = [ 1. 11.]

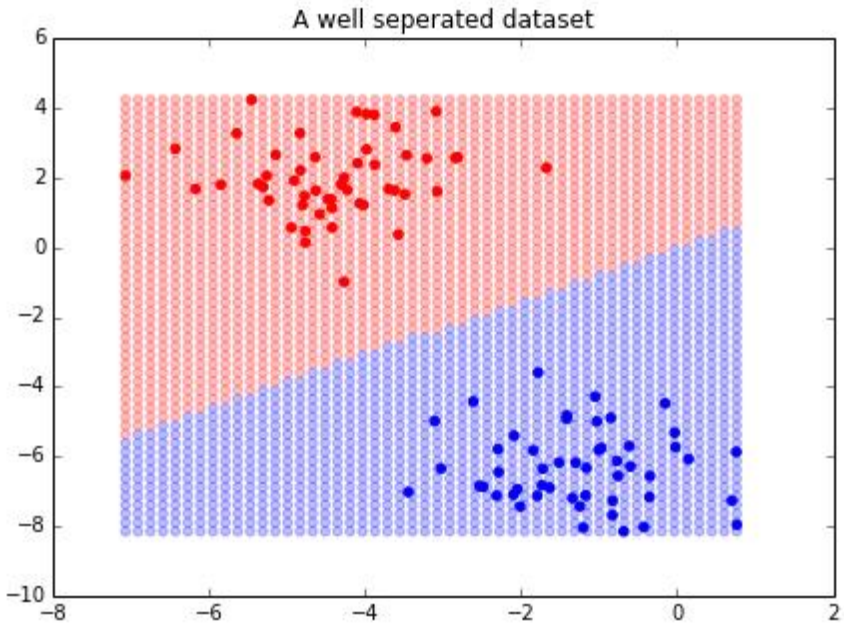

X[13] <= -0.0043
entropy = 0.999988458409
samples = 1000
X[14] <= -1.0793
entropy = 0.69483776707
samples = 578
X[18] <= -1.6464
entropy = 0.387310510123
samples = 422
entropy = 0.0000
samples = 25
value = [ 0. 25.]
X[11] <= -0.3879
entropy = 0.610066524326
samples = 553
X[11] <= -0.7355
entropy = 0.276195427648
samples = 21
X[13] <= 1.0368
entropy = 0.194016535661
samples = 401
X[14] <= 1.0360
entropy = 0.347428389888
samples = 430
X[6] <= -0.3401
entropy = 0.991927046405
samples = 123
entropy = 0.0000
samples = 1
value = [ 0. 1.]
entropy = 0.0000
samples = 20
value = [ 20. 0.]
X[18] <= 2.6845
entropy = 0.371232326641
samples = 154
X[2] <= -2.2244
entropy = 0.0380086554169
samples = 247
X[19] <= 1.5617
entropy = 0.0419125951035
samples = 220
X[6] <= -0.6707
entropy = 0.553507265455
samples = 210
X[5] <= 0.5058
entropy = 0.972359464368
samples = 82
X[16] <= 1.1175
entropy = 0.600508575413
samples = 41
X[6] <= 1.0047
entropy = 0.324302999152
samples = 152
entropy = 0.0000
samples = 2
value = [ 2. 0.]
X[3] <= -0.8764
entropy = 0.918295834054
samples = 3
entropy = 0.0000
samples = 244
value = [ 0. 244.]
entropy = 0.0000
samples = 209
value = [ 209. 0.]
entropy = 0.4395
samples = 11
value = [ 10. 1.]
entropy = 0.6827
samples = 149
value = [ 122. 27.]
entropy = 0.0000
samples = 61
value = [ 61. 0.]
entropy = 0.7950
samples = 50
value = [ 12. 38.]
entropy = 0.9284
samples = 32
value = [ 21. 11.]
entropy = 0.3160
samples = 35
value = [ 33. 2.]
entropy = 0.9183
samples = 6
value = [ 2. 4.]
entropy = 0.4365
samples = 100
value = [ 9. 91.]
entropy = 0.0000
samples = 52
value = [ 0. 52.]
entropy = 0.0000
samples = 1
value = [ 1. 0.]
entropy = 0.0000
samples = 2
value = [ 0. 2.]

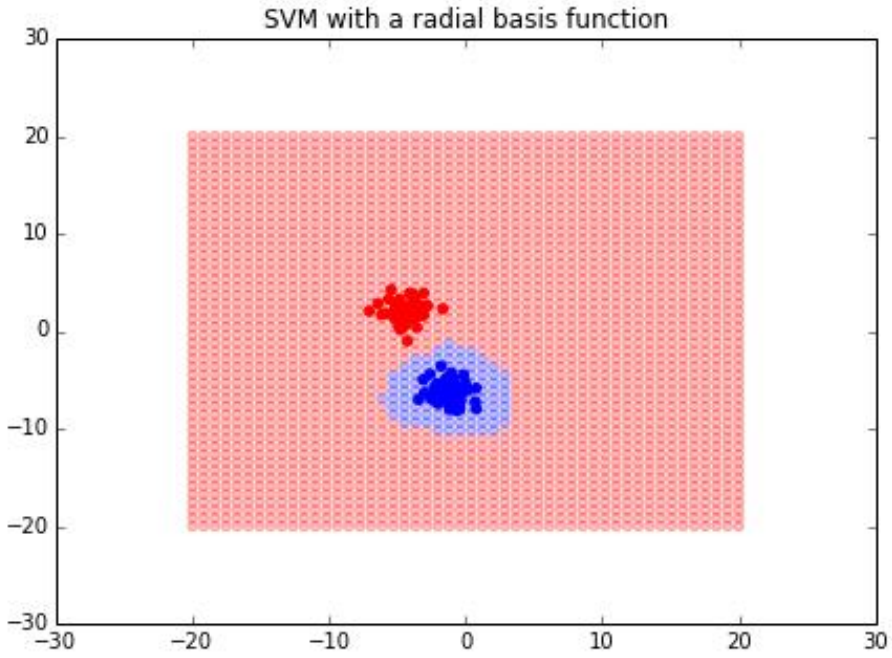

Accuracy at 0 class probability
% Correct
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
% trees for 0

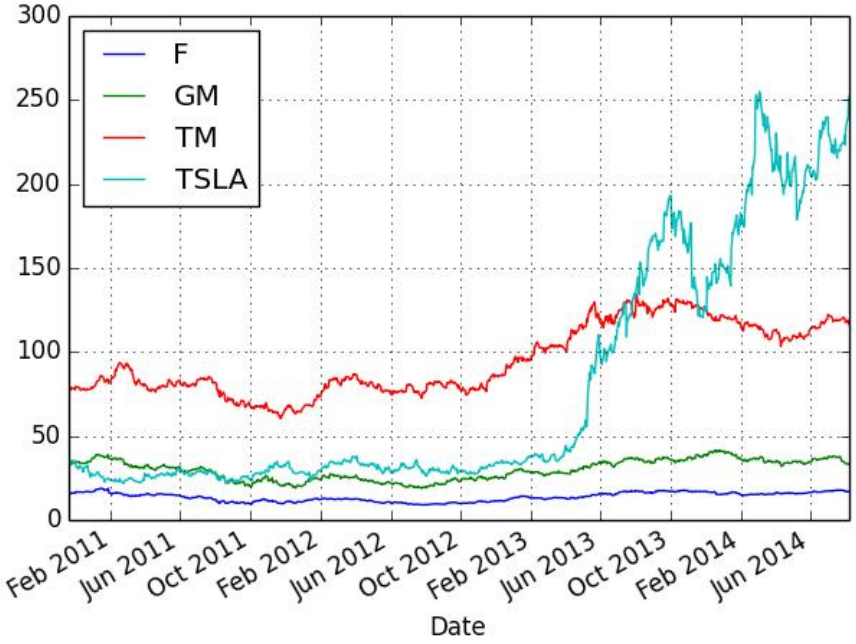

Guessed vs Correct (i, j) where i is the guess and j is the actual.
0.01
0.5
0.99
auto
log2
sqrt
Correct
1000
800
600
400
200
0
(0, 0)
(0, 1)
(1, 0)
(1, 1)
Guessed vs Correct

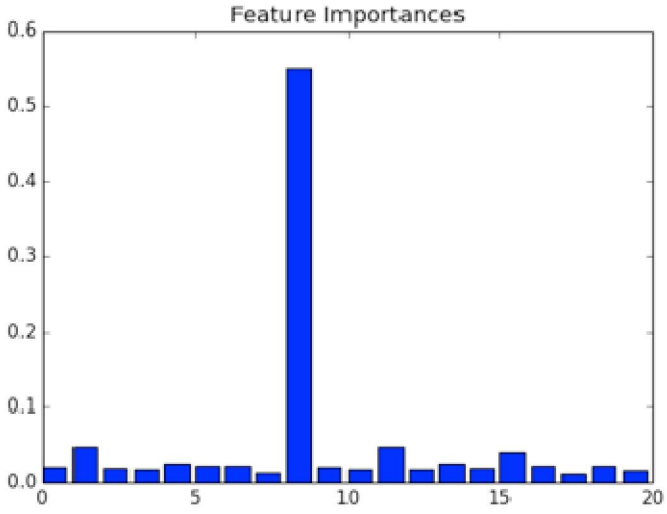

Accuracy by Number of Estimators
Accuracy
Number of Estimators
1.0
0.8
0.6
0.4
0.2
0.0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19


A well seperated dataset
4
2
0
−2
−4
−6
−8
−7
−6
−5
−4
−3
−2
−1
0


A well seperated dataset
6
4
2
0
−2
−4
−6
−8
−10
−8
−6
−4
−2
0
2


SVM with a radial basis function
30
20
10
0
−10
−20
−30
−30
−20
−10
0
10
20
30

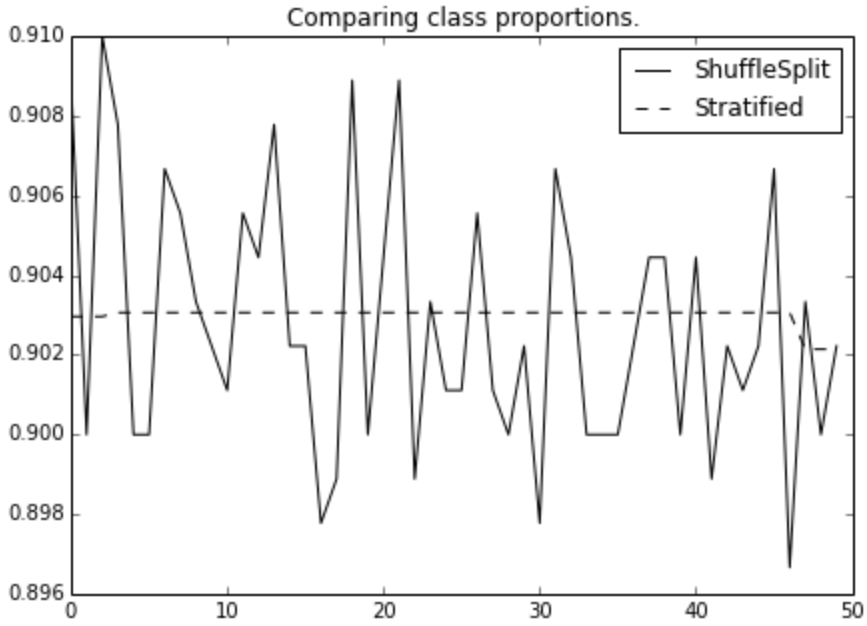

F
GM
TM
TSLA
300
250
200
150
100
50
0
Feb 2011
Jun 2011
Oct 2011
Feb 2012
Jun 2012
Oct 2012
Feb 2013
Jun 2013
Oct 2013
Feb 2014
Jun 2014
Date

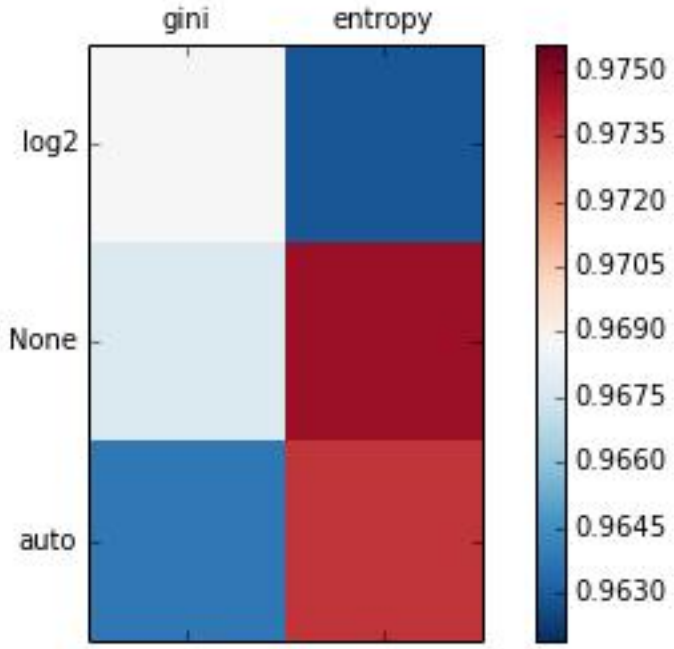

Feature Importances
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
5
10
15
20

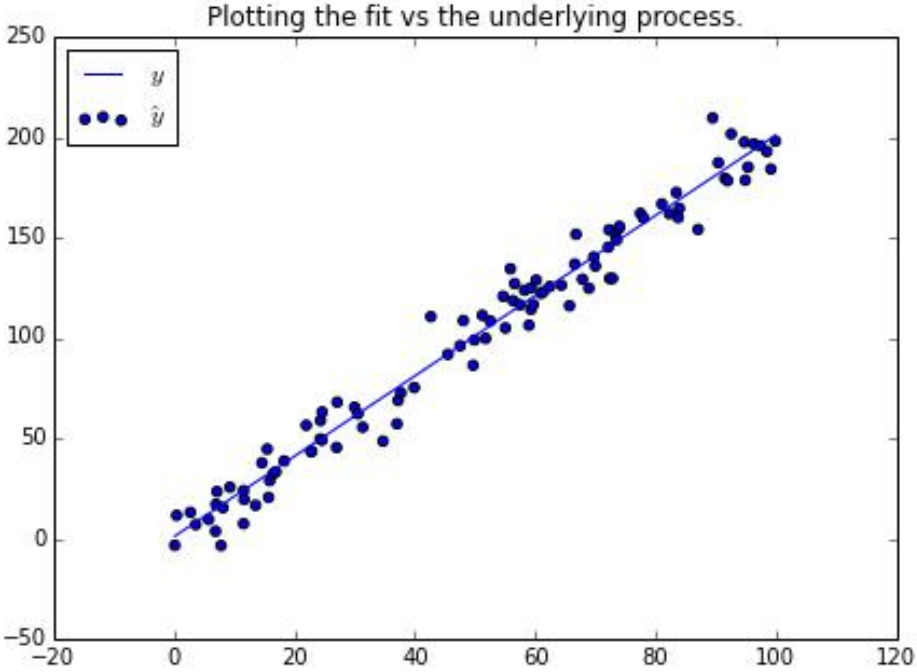
| | | Open | High | Low | Close | Volume | Adj Close | is_higher |
|---|---|---|---|---|---|---|---|---|
| **Date** | **minor** | | | | | | | |
| **2010-11-18** | **F** | 16.77 | 16.87 | 16.05 | 16.12 | 256937900 | 15.07 | 0 |
| | **GM** | 35.00 | 35.99 | 33.89 | 34.19 | 457044300 | 33.61 | 0 |
| | **TM** | 77.36 | 77.51 | 76.83 | 77.29 | 989100 | 77.29 | 0 |
| | **TSLA** | 30.67 | 30.74 | 28.92 | 29.89 | 956100 | 29.89 | 0 |
| **2010-11-19** | **F** | 16.02 | 16.38 | 15.83 | 16.28 | 130323600 | 15.22 | 0 |

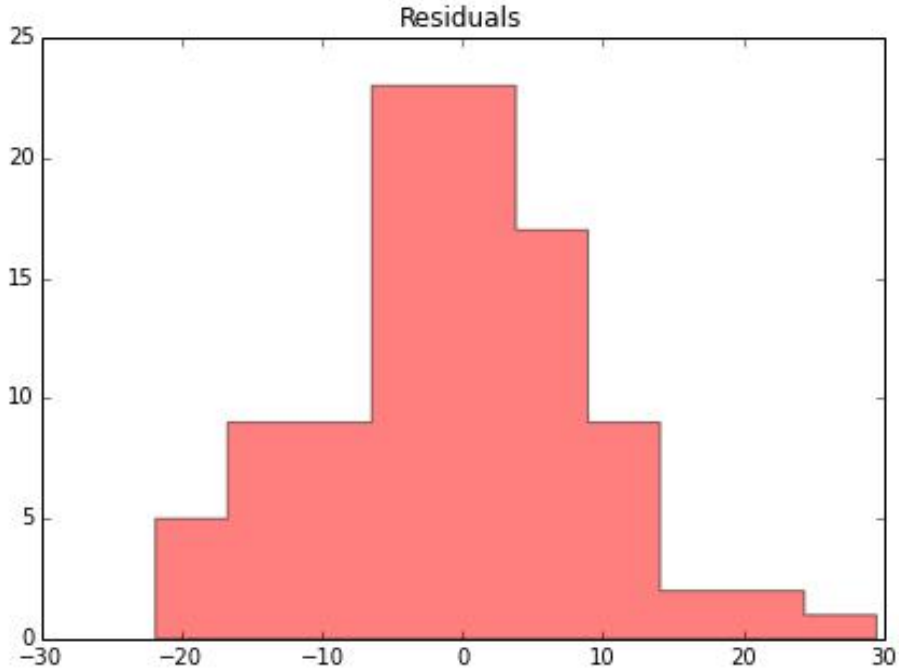
| | Open | High | Low | Close | Volume | is_higher |
|---|---|---|---|---|---|---|
| **0** | 16.77 | 16.87 | 16.05 | 16.12 | 256937900 | 0 |
| **1** | 35.00 | 35.99 | 33.89 | 34.19 | 457044300 | 0 |
| **2** | 77.36 | 77.51 | 76.83 | 77.29 | 989100 | 0 |
| **3** | 30.67 | 30.74 | 28.92 | 29.89 | 956100 | 0 |
| **4** | 16.02 | 16.38 | 15.83 | 16.28 | 130323600 | 0 |

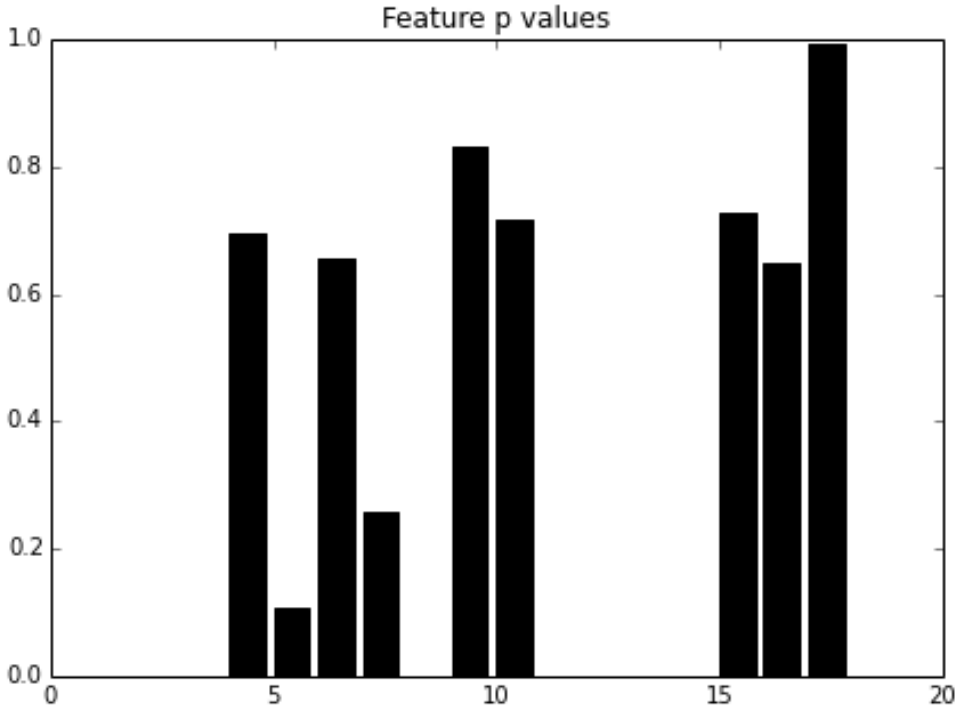
# Chapter 5

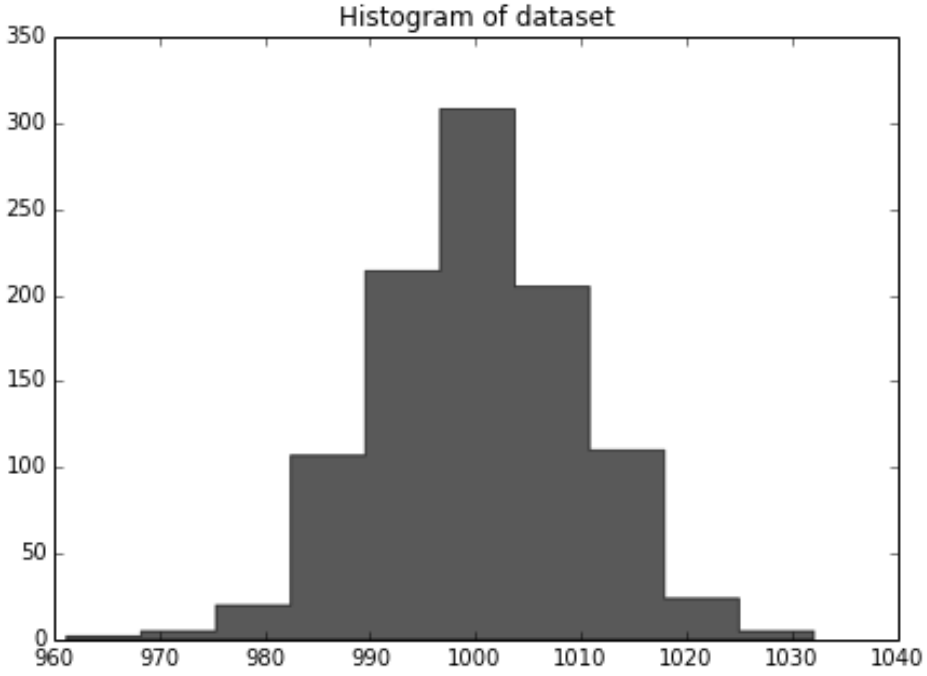

Comparing class proportions.
ShuffleSplit
Stratified
0.910
0.908
0.906
0.904
0.902
0.900
0.898
0.896
0
10
20
30
40
50

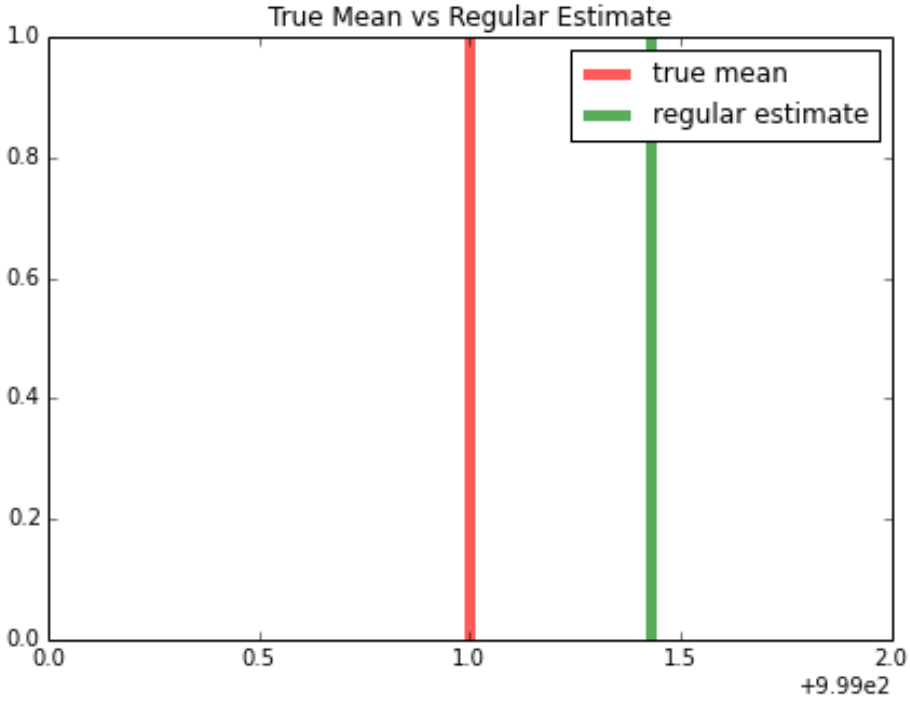

gini
entropy
log2
None
auto
0.9750
0.9735
0.9720
0.9705
0.9690
0.9675
0.9660
0.9645
0.9630

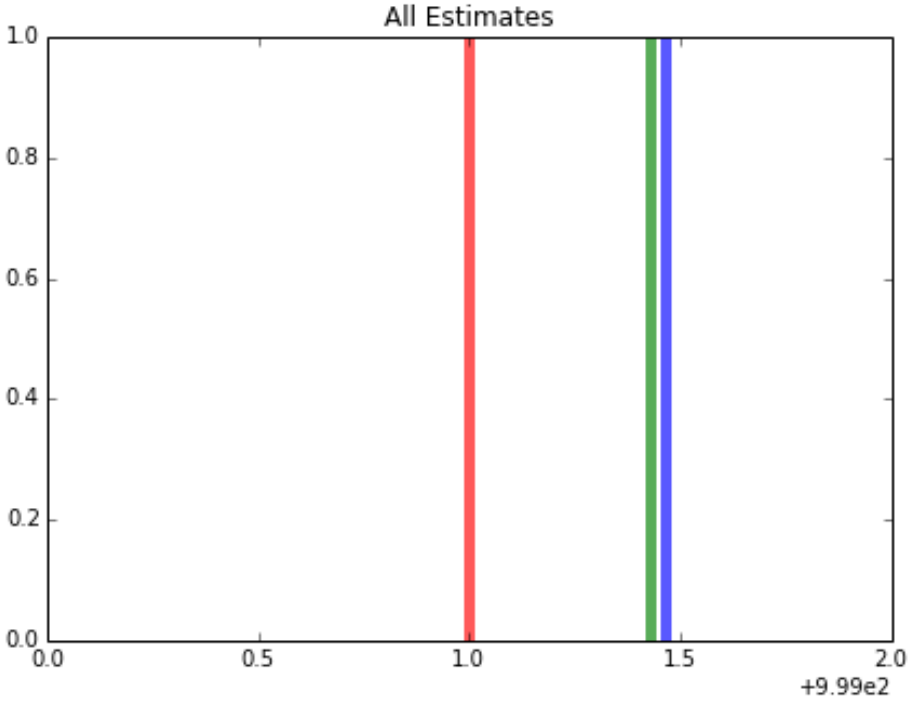

Plotting the fit vs the underlying process.
$y$
$\hat{y}$
250
200
150
100
50
0
−50
−20
0
20
40
60
80
100
120


Residuals
25
20
15
10
5
0
−30
−20
−10
0
10
20
30


Feature p values
1.0
0.8
0.6
0.4
0.2
0.0
0
5
10
15
20


Histogram of dataset
350
300
250
200
150
100
50
0
960
970
980
990
1000
1010
1020
1030
1040


True Mean vs Regular Estimate
true mean
regular estimate
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.5
1.0
1.5
2.0
+9.99e2


All Estimates
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.5
1.0
1.5
2.0
+9.99e2