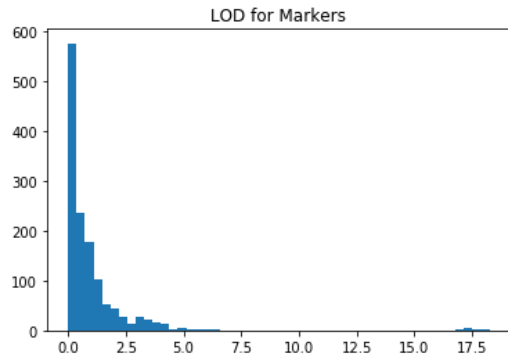


CSE 527 HW 2

Chenchao Xu

1. Quantitative trait loci analysis for cholesterol levels (code_q1.py)

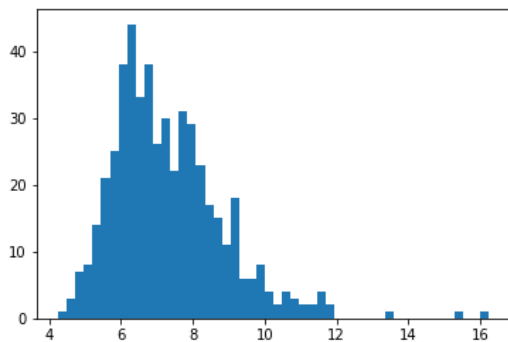
a.



Max LOD Score: 18.26

Number(index starts from 0) of the marker with Max LOD Score: 846

b.



95th percentile threshold = 10.015 (permutation_time = 500)

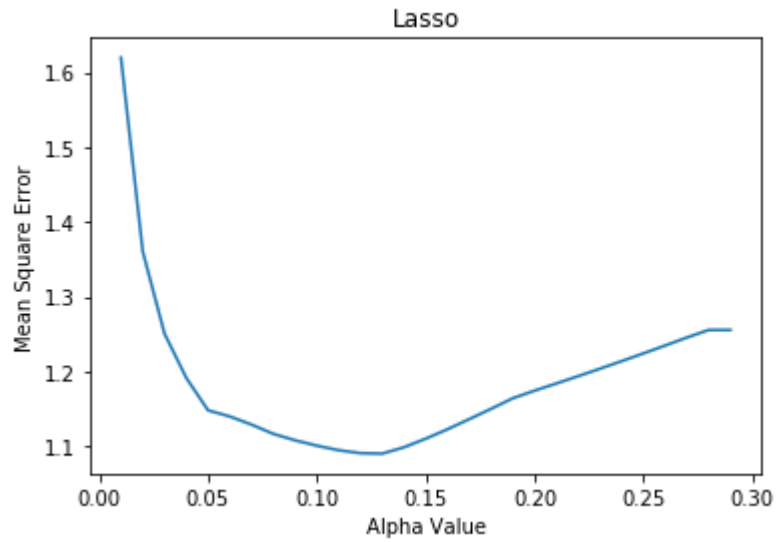
c.

[43, 124, 273, 561, 754, 846, 1123]

index starts form 0

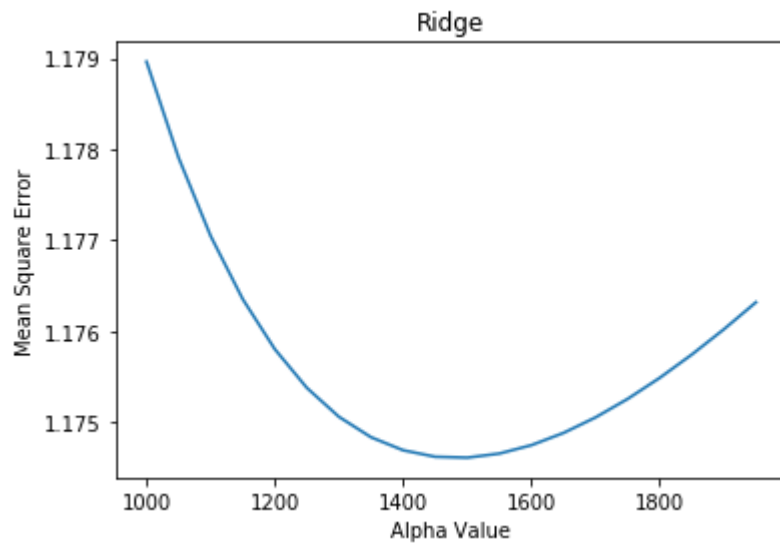
2. Multi-marker model for QTL analysis (code_q2.py)

a.



$\min(\text{lasso_mse}) = 1.090$

b.



$\min(\text{ridge_mse}) = 1.175$

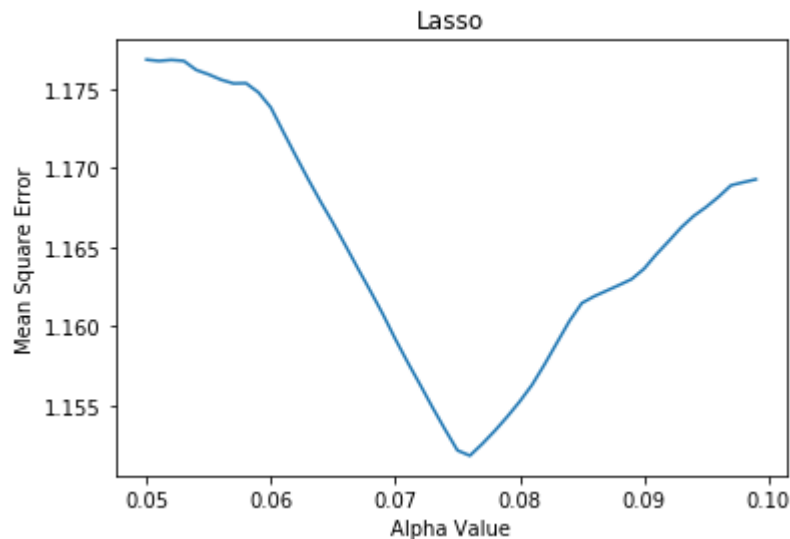
MSE with L2 is much higher than it with L1.

c.

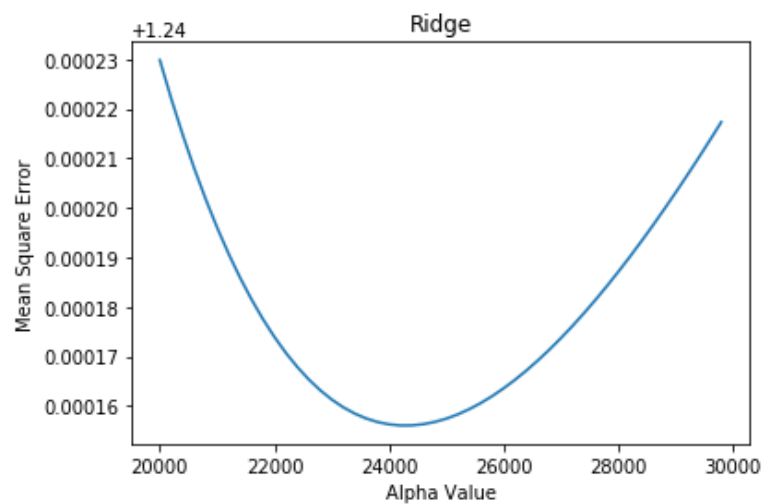
I tried to do leave-one-out on 250 training data for parameter tuning, it took too long to run on my laptop. So I used leave-one-set-out instead.

train set : validation set = 2:1; randomly splitted

And then use both training set and validation set to train the model for test set.



$\alpha_{\text{lasso}} = 0.076$



$\alpha_{\text{ridge}} = 24200$

On test set:

$\text{lasso_mse_test} = 1.120$; $\text{ridge_mse_test} = 1.234$; Lasso does better

parameters:

lasso: 0.076, not so far; ridge: 24200, much bigger than that in b

3. Implementation of the EM-based haplotype reconstruction algorithm (code_q3.py)

a.

hidden variables: haplotype states of individuals

parameters: haplotype frequencies

b.

Use current frequency estimates to replace ambiguous genotypes with fractional counts of phased genotypes

Probability Table:

0.4 , 0.6
0.75 , 0.25
0.6 , 0.4

c.

Estimate frequency of each haplotype by counting

Frequency Table:

```
[ 0.125      0.04166667  0.06666667  0.04166667  0.325      0.1
  0.13333333  0.06666667  0.1          ]
```

d.

iteration_time: 6

Probability Table:

9.46389217332e-25 , 1.0
1.0 , 2.4739168422e-58
1.0 , 9.46389217332e-25

Frequency Table:

```
[ 1.66666667e-01  4.12319474e-59  1.57731536e-25
 4.12319474e-59
 5.00000000e-01  1.66666667e-01  3.15463072e-25
 1.57731536e-25
 1.66666667e-01]
```

covered!!!

The last iteration_time is 6.

Appendix:

Results for code 3

iteration_time: 0

Probability Table:

0.4 , 0.6

0.75 , 0.25

0.6 , 0.4

Frequency Table:

[0.125 0.04166667 0.06666667 0.04166667 0.325 0.1
0.13333333 0.06666667 0.1]

iteration_time: 1

Probability Table:

0.214765100671 , 0.785234899329

0.959016393443 , 0.0409836065574

0.785234899329 , 0.214765100671

Frequency Table:

[0.15983607 0.0068306 0.03579418 0.0068306 0.42158103 0.13087248
0.07158837 0.03579418 0.13087248]

iteration_time: 2

Probability Table:

0.0443822892185 , 0.955617710782

0.999308071215 , 0.000691928785389

0.955617710782 , 0.0443822892185

Frequency Table:

[1.66551345e-01 1.15321464e-04 7.39704820e-03 1.15321464e-04
4.85090582e-01 1.59269618e-01 1.47940964e-02 7.39704820e-03

1.59269618e-01]

iteration_time: 3

Probability Table:

0.00141441353543 , 0.998585586465

0.999999835393 , 1.64607355067e-07

0.998585586465 , 0.00141441353543

Frequency Table:

[1.66666639e-01 2.74345592e-08 2.35735589e-04 2.74345592e-08

4.99528501e-01 1.66430931e-01 4.71471178e-04 2.35735589e-04

1.66430931e-01]

iteration_time: 4

Probability Table:

1.33685839109e-06 , 0.999998663142

1.0 , 9.04038699399e-15

0.999998663142 , 1.33685839109e-06

Frequency Table:

[1.66666667e-01 1.50673117e-15 2.22809732e-07 1.50673117e-15

4.99999554e-01 1.66666444e-01 4.45619464e-07 2.22809732e-07

1.66666444e-01]

iteration_time: 5

Probability Table:

1.19146289325e-12 , 0.999999999999

1.0 , 2.72428899469e-29

0.999999999999 , 1.19146289325e-12

Frequency Table:

```
[ 1.66666667e-01  4.54048166e-30  1.98577149e-13  4.54048166e-30
 5.00000000e-01  1.66666667e-01  3.97154298e-13  1.98577149e-13
 1.66666667e-01]
```

iteration_time: 6

Probability Table:

9.46389217332e-25 , 1.0

1.0 , 2.4739168422e-58

1.0 , 9.46389217332e-25

Frequency Table:

```
[ 1.66666667e-01  4.12319474e-59  1.57731536e-25  4.12319474e-59
 5.00000000e-01  1.66666667e-01  3.15463072e-25  1.57731536e-25
 1.66666667e-01]
```

covered!!!