Pset 5
1a.

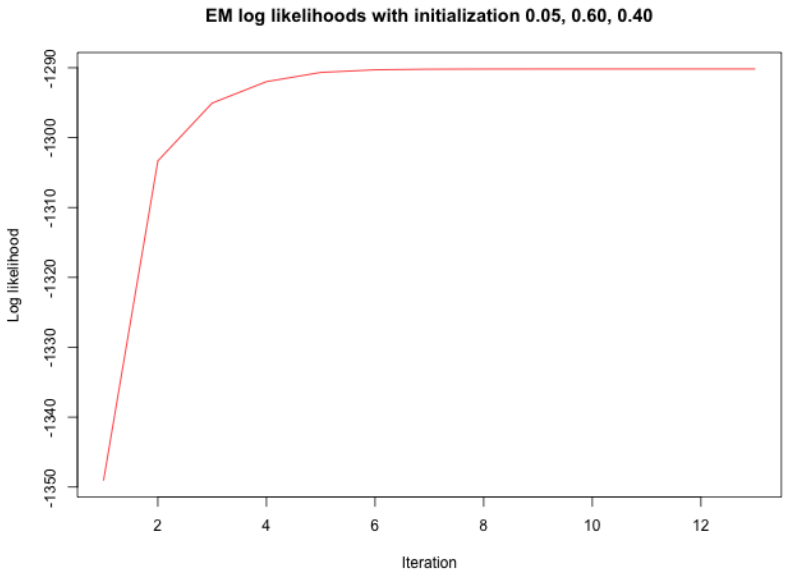
Note:

Mu = mean(c(a_kt['h','l'],a_kt['l','h']))

theta_h: 0.78140

theta_l: 0.34561

mu: 0.00861



note: stop_diff = 0.0001

1b:

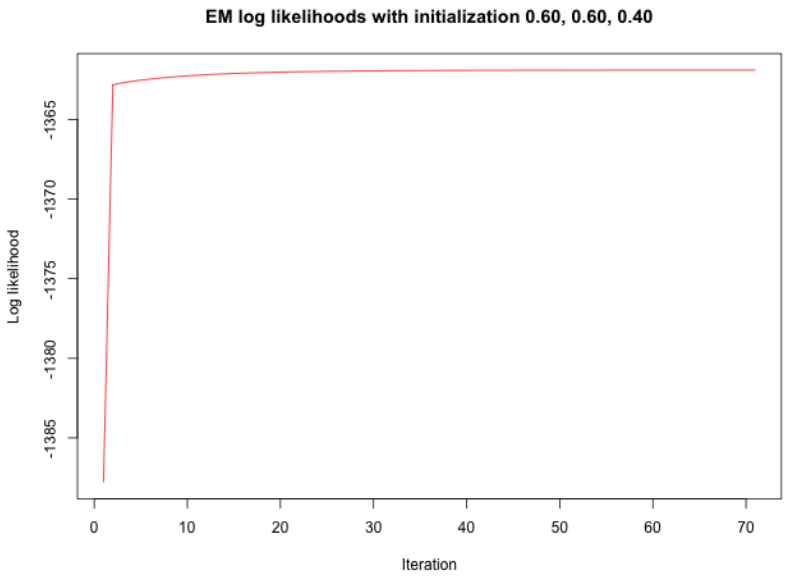
C(mu,theta_h,theta_l) = 0.6,0.6,0.4

Output:

theta_h: 0.61359

theta_l: 0.60763

mu: 0.57515

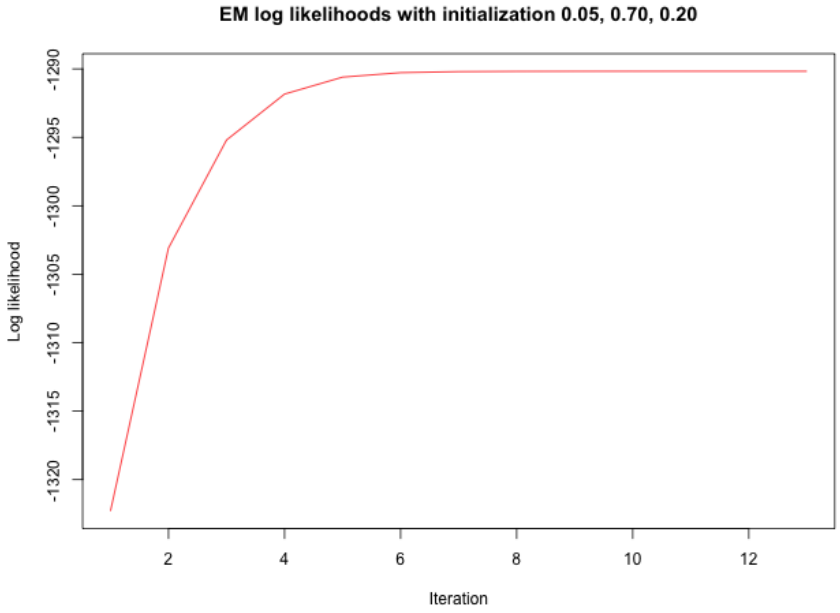


C(mu,theta_h,theta_l) = 0.05,0.7,0.2

theta_h: 0.78140

theta_l: 0.34561

mu: 0.00861

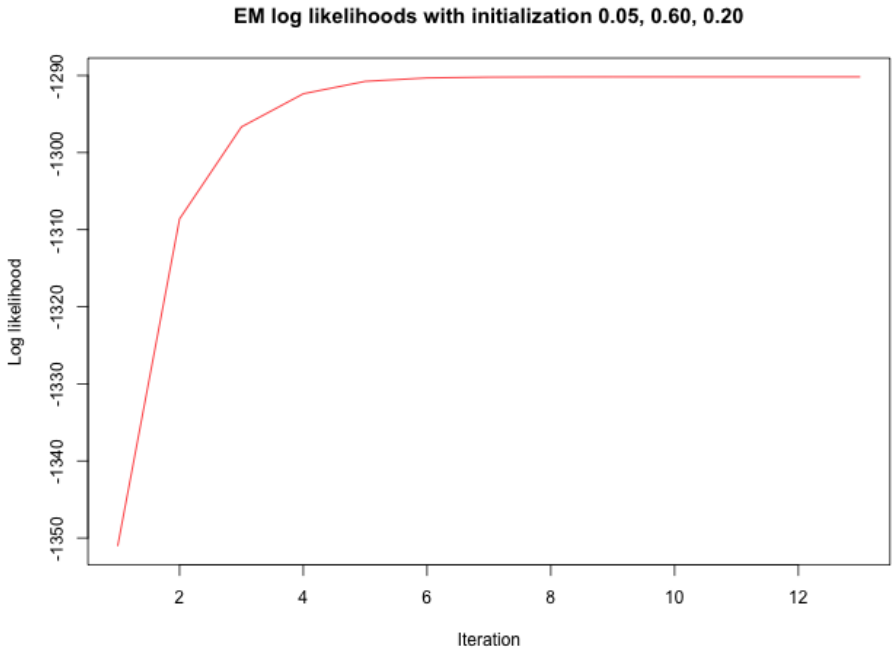


C(mu,theta_h,theta_l) = 0.05,0.6,0.2

theta_h: 0.78140

theta_l: 0.34561

mu: 0.00861



When mu>0.5, the EM is stuck in local maximum. (mu>0.5 is not expected biologically)
Multiple tests are required to output the universal optimum.
When the initial parameters are reasonable, the EM output the optimum with a few more iterations.

2a.

#vectorization with mapply

#by recruiting factor matrix

Termination criteria:

iteration >5000 | sum(abs(pre_pwm-pwm))<exp(-5)

#(when the difference between is smaller than exp(-5)

Epsilon = 1

2b.

Motif found after multiple tests:

CTCACTGGAC

Tryout 1:

Sequence 1: TCACTGGACC

Sequence 2: TCACTGGACT

Sequence 3: TCACTGGACT

Sequence 4: TCACTGGACG

Sequence 5: TCACTGGACT

Sequence 6: TCACTGGACC

Sequence 7: TCACTGGACG

Sequence 8: TCACTGGACC

Sequence 9: TCACTGGACC

Sequence 10: TCACTGGACT

Majority motif (by wj): TCACTGGACC

	A	T	C	G
[1,]	0.07692308	0.76923077	0.07692308	0.07692308
[2,]	0.07692308	0.07692308	0.76923077	0.07692308
[3,]	0.76923077	0.07692308	0.07692308	0.07692308
[4,]	0.07692308	0.07692308	0.76923077	0.07692308
[5,]	0.07692308	0.76923077	0.07692308	0.07692308
[6,]	0.07692308	0.07692308	0.07692308	0.76923077

[7,] 0.07692308 0.07692308 0.07692308 0.76923077

[8,] 0.76923077 0.07692308 0.07692308 0.07692308

[9,] 0.07692308 0.07692308 0.76923077 0.07692308

[10,] 0.07692308 0.30769231 0.38461538 0.23076923

Tryout2:

Sequence 1: CTCACTGGAC

Sequence 2: CTCACTGGAC

Sequence 3: CTCACTGGAC

Sequence 4: CTCACTGGAC

Sequence 5: CTCACTGGAC

Sequence 6: CTCACTGGAC

Sequence 7: CTCACTGGAC

Sequence 8: CTCACTGGAC

Sequence 9: CTCACTGGAC

Sequence 10: CTCACTGGAC

Majority motif (by wj): CTCACTGGAC

Tryout3:

Sequence 1: CTCACTGGAC

Sequence 2: CTCACTGGAC

Sequence 3: CTCACTGGAC

Sequence 4: CTCACTGGAC

Sequence 5: CTCACTGGAC

Sequence 6: CTCACTGGAC

Sequence 7: CTCACTGGAC

Sequence 8: CTCACTGGAC

Sequence 9: CTCACTGGAC

Sequence 10: CTCACTGGAC

Majority motif (by wj): CTCACTGGAC

Tryout 4:

Sequence 1: GTCTCACTGG

Sequence 2: CATTCACTGG

Sequence 3: GCCTCACTGG

Sequence 4: GGCTCACTGG

Sequence 5: TGCTCACTGG

Sequence 6: TACTCACTGG

Sequence 7: TTCTCACTGG

Sequence 8: TTCTCACTGG

Sequence 9: CTCTCACTGG

Sequence 10: GCCTCACTGG

Majority motif (by wj): TTCTCACTGG

	A	T	C	G
[1,]	0.07692308	0.38461538	0.23076923	0.30769231
[2,]	0.23076923	0.38461538	0.15384615	0.23076923
[3,]	0.07692308	0.15384615	0.69230769	0.07692308
[4,]	0.07692308	0.76923077	0.07692308	0.07692308
[5,]	0.07692308	0.07692308	0.76923077	0.07692308
[6,]	0.76923077	0.07692308	0.07692308	0.07692308
[7,]	0.07692308	0.07692308	0.76923077	0.07692308
[8,]	0.07692308	0.76923077	0.07692308	0.07692308
[9,]	0.07692308	0.07692308	0.07692308	0.76923077
[10,]	0.07692308	0.07692308	0.07692308	0.76923077

2c.

Summary:

Motif list:

Test:	Motif reported:

1	CATGAACCAT
2	CCCCTGTGGG
3	TGCAGACGGA
4	ATTGAATTAT
5	GGCACAAGCT
6	CTTTCAGGAC
7	ATCTGACATT

2nd is more likely to be the motif.

Change criteria to #iteration<500|sum(abs(diff_wj))<exp(-2)

Test1:

Sequence 1: GATAATCCAT

Sequence 2: AGTGCTCCCC

Sequence 3: TATGCTACAC

Sequence 4: CATGTATCTT

Sequence 5: GATACAGCCT

Sequence 6: GTTCAAACTT

Sequence 7: CTGGACCCGT

Sequence 8: CATAAACCAT

Sequence 9: CATGCTCCAT

Sequence 10: TGGTCCCAGT

Majority motif (by wj): CATGAACCAT

	A	T	C	G
[1,]	0.15384615	0.15384615	0.38461538	0.30769231
[2,]	0.53846154	0.23076923	0.07692308	0.15384615
[3,]	0.07692308	0.69230769	0.07692308	0.15384615
[4,]	0.30769231	0.07692308	0.15384615	0.46153846
[5,]	0.38461538	0.15384615	0.38461538	0.07692308
[6,]	0.38461538	0.38461538	0.15384615	0.07692308
[7,]	0.23076923	0.15384615	0.46153846	0.15384615

[8,] 0.07692308 0.07692308 0.76923077 0.07692308

[9,] 0.38461538 0.23076923 0.23076923 0.15384615

[10,] 0.07692308 0.61538462 0.23076923 0.07692308

Test2:

Sequence 1: CCCCATCGGG

Sequence 2: CACCTGTGGA

Sequence 3: CCCCCGAGGT

Sequence 4: CCCGTATTGG

Sequence 5: CCTCTGAGGA

Sequence 6: CCCCAATGCT

Sequence 7: CCCCATTCTG

Sequence 8: GACGTATCAT

Sequence 9: GACCTGTCGG

Sequence 10: CCTGTGTGCA

Majority motif (by wj): CCCCTGTGGG

	A	T	C	G
[1,]	0.07692308	0.07692308	0.61538462	0.23076923
[2,]	0.30769231	0.07692308	0.53846154	0.07692308
[3,]	0.07692308	0.15384615	0.69230769	0.07692308
[4,]	0.07692308	0.07692308	0.61538462	0.23076923
[5,]	0.30769231	0.46153846	0.15384615	0.07692308
[6,]	0.30769231	0.23076923	0.07692308	0.38461538
[7,]	0.23076923	0.53846154	0.15384615	0.07692308
[8,]	0.07692308	0.15384615	0.30769231	0.46153846
[9,]	0.15384615	0.15384615	0.15384615	0.53846154
[10,]	0.23076923	0.30769231	0.07692308	0.38461538

Test3:

Sequence 1: CGCCGACCGA

Sequence 2: TGGGCAGGGC

Sequence 3: CGTAGAGGGC

Sequence 4: GGAAATCGGC

Sequence 5: TACAGACGAA

Sequence 6: GGATATCAGT

Sequence 7: TGCGGAGAAA

Sequence 8: GTATGCGGCA

Sequence 9: AGCACACTTC

Sequence 10: GGTACTCAGG

Majority motif (by wj): TGCAGACGGA

	A	T	C	G
[1,]	0.15384615	0.30769231	0.23076923	0.30769231
[2,]	0.15384615	0.15384615	0.07692308	0.61538462
[3,]	0.30769231	0.15384615	0.38461538	0.15384615
[4,]	0.38461538	0.23076923	0.15384615	0.23076923
[5,]	0.23076923	0.07692308	0.23076923	0.46153846
[6,]	0.53846154	0.23076923	0.15384615	0.07692308
[7,]	0.07692308	0.07692308	0.46153846	0.38461538
[8,]	0.23076923	0.15384615	0.15384615	0.46153846
[9,]	0.23076923	0.15384615	0.15384615	0.46153846
[10,]	0.38461538	0.15384615	0.38461538	0.07692308

Test 4:

Sequence 1: CATGAAATGT

Sequence 2: ATTGAACTAT

Sequence 3: ATTGTTATAG

Sequence 4: AAAGTCTCAT

Sequence 5: AACCAATTAA

Sequence 6: ATTGAATTTA

Sequence 7: ATTGATCTGA

Sequence 8: CTAAGTTTGG

Sequence 9: AGTCATATTT

Sequence 10: CGAGACCTAG

Majority motif (by wj): ATTGAATTAT

```
              A         T         C         G
 [1,] 0.61538462 0.07692308 0.23076923 0.07692308
 [2,] 0.30769231 0.46153846 0.07692308 0.15384615
 [3,] 0.23076923 0.53846154 0.15384615 0.07692308
 [4,] 0.15384615 0.07692308 0.23076923 0.53846154
 [5,] 0.53846154 0.23076923 0.07692308 0.15384615
 [6,] 0.38461538 0.38461538 0.15384615 0.07692308
 [7,] 0.30769231 0.38461538 0.23076923 0.07692308
 [8,] 0.07692308 0.69230769 0.15384615 0.07692308
 [9,] 0.38461538 0.23076923 0.07692308 0.30769231
[10,] 0.30769231 0.38461538 0.07692308 0.23076923
```

Test 5:

Sequence 1: GATACATGAT

Sequence 2: GACACATGAA

Sequence 3: GGGATATGCT

Sequence 4: GCGCCATCCA

Sequence 5: AGCCGAAACT

Sequence 6: AGCAGAAGTT

Sequence 7: GACAAAAGCA

Sequence 8: GTACTAAGTT

Sequence 9: AGCACAAGCT

Sequence 10: AGGCGGACCT

Majority motif (by wj): GGCACAAGCT

```
            A         T         C         G
 [1,] 0.3076923 0.07692308 0.07692308 0.53846154
```

[2,] 0.3076923 0.15384615 0.15384615 0.38461538

[3,] 0.1538462 0.15384615 0.46153846 0.23076923

[4,] 0.5384615 0.07692308 0.30769231 0.07692308

[5,] 0.1538462 0.23076923 0.38461538 0.23076923

[6,] 0.7692308 0.07692308 0.07692308 0.07692308

[7,] 0.4615385 0.38461538 0.07692308 0.07692308

[8,] 0.1538462 0.07692308 0.15384615 0.61538462

[9,] 0.2307692 0.23076923 0.46153846 0.07692308

[10,] 0.3076923 0.53846154 0.07692308 0.07692308

Test6:

Sequence 1: CTTTCGTGTC

Sequence 2: ATTTGATGAG

Sequence 3: CTTTTACCCA

Sequence 4: CTTTCGTTAC

Sequence 5: ATTTGAGTCC

Sequence 6: GTCTGAGAAA

Sequence 7: CTGTGCGCAC

Sequence 8: TCTTCGGGAG

Sequence 9: CGTTCGAGAC

Sequence 10: CTCTCTCCCC

Majority motif (by wj): CTTTCAGGAC

A T C G

[1,] 0.23076923 0.15384615 0.46153846 0.15384615

[2,] 0.07692308 0.61538462 0.15384615 0.15384615

[3,] 0.07692308 0.61538462 0.15384615 0.15384615

[4,] 0.07692308 0.76923077 0.07692308 0.07692308

[5,] 0.07692308 0.15384615 0.38461538 0.38461538

[6,] 0.38461538 0.07692308 0.15384615 0.38461538

[7,] 0.15384615 0.30769231 0.15384615 0.38461538

[8,] 0.15384615 0.23076923 0.23076923 0.38461538

[9,] 0.53846154 0.15384615 0.23076923 0.07692308

[10,] 0.23076923 0.07692308 0.46153846 0.23076923

Test7:

Sequence 1: TGATGACCAC

Sequence 2: ATCTACCATA

Sequence 3: AGCTGCAATG

Sequence 4: ACCCAATGTT

Sequence 5: CTTTGGCATT

Sequence 6: AGCAGAAGTT

Sequence 7: CTATCACACT

Sequence 8: CCCCCGAAAG

Sequence 9: CCCCTATAAG

Sequence 10: CCCTAACATT

Majority motif (by wj): ATCTGACATT

A T C G

[1,] 0.38461538 0.15384615 0.3846154 0.07692308

[2,] 0.07692308 0.30769231 0.3076923 0.30769231

[3,] 0.23076923 0.15384615 0.5384615 0.07692308

[4,] 0.15384615 0.46153846 0.3076923 0.07692308

[5,] 0.23076923 0.15384615 0.2307692 0.38461538

[6,] 0.46153846 0.07692308 0.2307692 0.23076923

[7,] 0.30769231 0.23076923 0.3846154 0.07692308

[8,] 0.53846154 0.07692308 0.1538462 0.23076923

[9,] 0.30769231 0.46153846 0.1538462 0.07692308

[10,] 0.15384615 0.38461538 0.1538462 0.30769231

2d. in motif1.fa, different runs yield similar results with one or two off targets.

It is hard to get consensus in motif2.fa

The stop criteria and the random sample method may influence the outcomes in each run. The characteristics of data will also interfere. If there are lots of confounding small repeats show up in the data, it may cover the real motif.