Homework 1

Song Jiecheng 13300180032

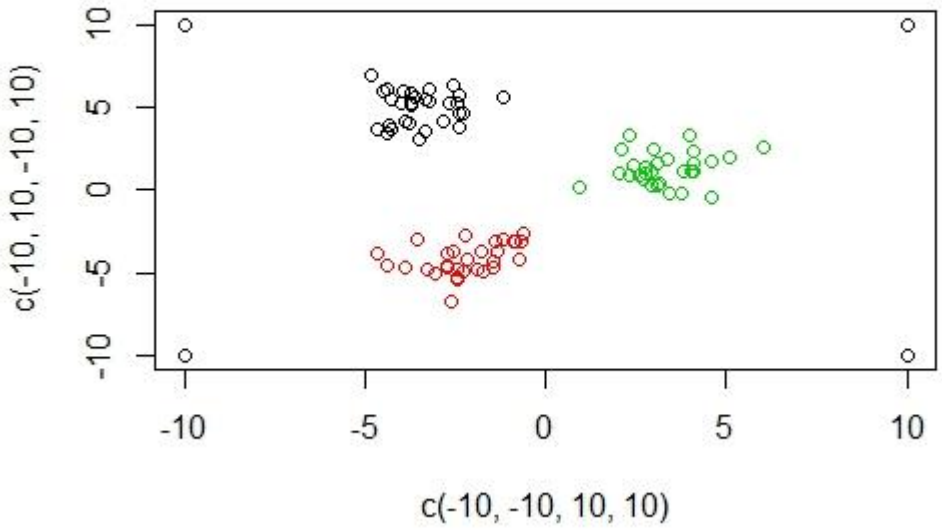
Question 1 : Rewrite the target function of K-means in matrix form.

$$f(x)= \operatorname{argmin}_s \sum_{a=1}^{k} \sum_{j \epsilon S_a} \sqrt{\sum_{i=1}^{m} (A_{ij} - C_{ia})^2}$$

A is the data, an m*n matrix, C_a is the center of the clusters.

Question 2 : Cluster Data1(2/3/4).csv with k-means, judging the number of clusters and compare the different evaluating methods.

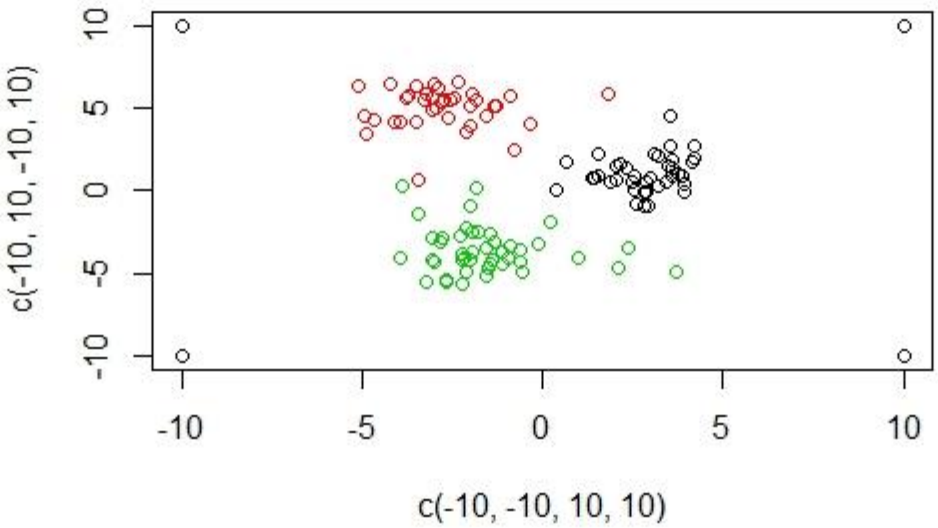
1. Data1.csv



Number of clusters: 3

Clustering Condition: 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2

2. Data2.csv



Number of clusters: 3

Clustering condition: 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

1 1 1 1 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 1 2 1 2 2 2 2 3 3 1 3 2 3 1 1 1 3 3 3 1 1 3 3 3 1 3
3 3 3 2

3. Data3.csv

Number of clusters: 3

Clustering condition: 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 3 3 3 3

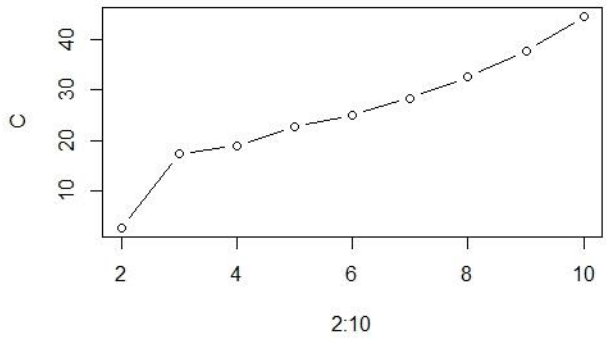
4. Data4.csv

Number of clusters: 3

Clustering condition: 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

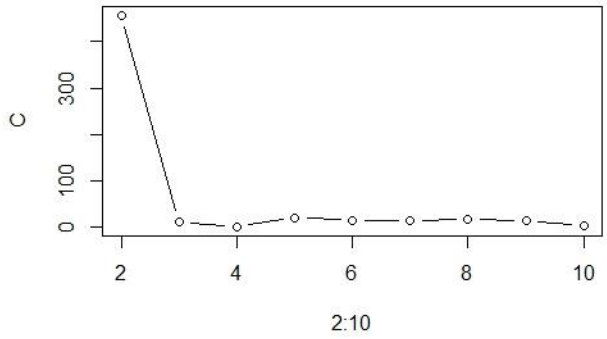
1 1 1 1 1 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 3 3 3 3 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
2 2 2 2 2 1 3 1 1 1 2 2 1 3 1 1 1 1 1 1 1 2 1 1 3 3 1 1 1 1 1
1 3 1 1

5. Calinski & Harabasz



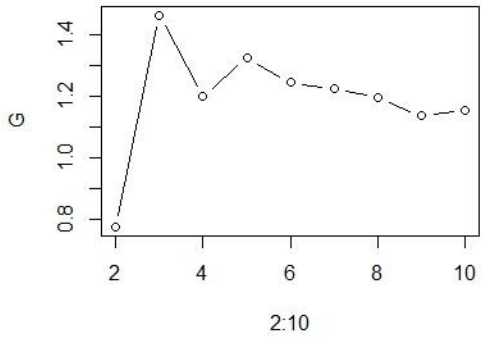
The picture is the result of C&H evaluating Data1

6. Hartigan



The picture is the result of C&H evaluating Data1

7. Gap Statistic



The picture is shows the 'gap' by using gap statistic evaluating Data1

8. Conclusion

After testing many times, I find that C&H is the most stable, and Hartigan and Gap Statistic sometimes are not as stable as C&H

Question 3: Use hierarchical clustering methods to cluster Data1/2.csv

1. Data1.csv

simple	complex	average
[,1] [,2]	[,1] [,2]	[,1] [,2]
[1,] 46 44	[1,] 46 44	[1,] 46 44
[2,] 51 31	[2,] 51 31	[2,] 51 31
[3,] 30 7	[3,] 30 7	[3,] 30 7
[4,] 6 5	[4,] 6 5	[4,] 6 5
[5,] 10 5	[5,] 90 74	[5,] 10 5
[6,] 90 74	[6,] 80 79	[6,] 90 74
[7,] 80 79	[7,] 89 73	[7,] 80 79
[8,] 89 73	[8,] 83 69	[8,] 89 73
[9,] 83 69	[9,] 58 43	[9,] 83 69
[10,] 58 43	[10,] 22 18	[10,] 58 43
[11,] 22 18	[11,] 49 47	[11,] 22 18
[12,] 49 47	[12,] 10 5	[12,] 49 47
[13,] 44 39	[13,] 71 70	[13,] 71 70
[14,] 71 70	[14,] 44 39	[14,] 11 5
[15,] 11 5	[15,] 52 36	[15,] 44 39
[16,] 21 18	[16,] 41 33	[16,] 52 36
[17,] 52 36	[17,] 11 5	[17,] 18 17
[18,] 18 17	[18,] 78 63	[18,] 28 7
[19,] 28 7	[19,] 86 79	[19,] 86 79

[20,]	86	79	[20,]	50	32	[20,]	41	33
[21,]	41	33	[21,]	68	65	[21,]	78	63
[22,]	78	63	[22,]	17	2	[22,]	50	32
[23,]	50	32	[23,]	28	7	[23,]	68	65
[24,]	68	65	[24,]	59	57	[24,]	17	2
[25,]	81	63	[25,]	29	8	[25,]	59	57
[26,]	17	2	[26,]	56	40	[26,]	29	8
[27,]	73	63	[27,]	66	62	[27,]	56	40
[28,]	8	2	[28,]	21	18	[28,]	66	62
[29,]	59	57	[29,]	27	26	[29,]	16	2
[30,]	29	2	[30,]	25	24	[30,]	27	26
[31,]	56	40	[31,]	81	73	[31,]	81	79
[32,]	66	62	[32,]	88	62	[32,]	25	24
[33,]	39	32	[33,]	55	53	[33,]	21	8
[34,]	16	2	[34,]	23	1	[34,]	36	33
[35,]	47	33	[35,]	82	72	[35,]	88	62
[36,]	27	26	[36,]	54	45	[36,]	39	35
[37,]	64	62	[37,]	87	84	[37,]	55	53
[38,]	79	63	[38,]	39	35	[38,]	33	31
[39,]	25	24	[39,]	16	2	[39,]	77	70
[40,]	33	31	[40,]	77	70	[40,]	26	2
[41,]	36	31	[41,]	57	47	[41,]	23	1
[42,]	88	62	[42,]	64	62	[42,]	82	72
[43,]	77	63	[43,]	75	67	[43,]	79	73
[44,]	5	2	[44,]	33	31	[44,]	54	45
[45,]	35	32	[45,]	43	34	[45,]	87	84
[46,]	57	31	[46,]	40	36	[46,]	70	63
[47,]	55	53	[47,]	18	8	[47,]	43	34
[48,]	70	63	[48,]	69	65	[48,]	73	65
[49,]	26	2	[49,]	76	73	[49,]	57	47
[50,]	40	31	[50,]	20	1	[50,]	64	62
[51,]	23	1	[51,]	70	63	[51,]	74	62
[52,]	82	72	[52,]	13	3	[52,]	9	1
[53,]	74	62	[53,]	26	14	[53,]	75	67
[54,]	67	65	[54,]	8	2	[54,]	67	65
[55,]	54	45	[55,]	35	32	[55,]	53	40
[56,]	87	84	[56,]	74	62	[56,]	35	32
[57,]	7	1	[57,]	47	31	[57,]	8	2
[58,]	76	63	[58,]	24	5	[58,]	47	34
[59,]	43	34	[59,]	9	4	[59,]	69	65
[60,]	65	63	[60,]	45	42	[60,]	60	40
[61,]	24	2	[61,]	60	53	[61,]	45	42
[62,]	9	1	[62,]	79	63	[62,]	40	34
[63,]	53	34	[63,]	7	1	[63,]	76	63

[64,] 75 63	[64,] 73 67	[64,] 20 1
[65,] 69 63	[65,] 84 72	[65,] 13 3
[66,] 34 31	[66,] 53 36	[66,] 7 1
[67,] 45 31	[67,] 48 38	[67,] 24 19
[68,] 84 63	[68,] 85 65	[68,] 14 3
[69,] 32 31	[69,] 34 31	[69,] 48 32
[70,] 20 1	[70,] 14 1	[70,] 5 2
[71,] 72 62	[71,] 19 5	[71,] 72 62
[72,] 60 31	[72,] 12 3	[72,] 34 31
[73,] 2 1	[73,] 63 61	[73,] 63 61
[74,] 42 31	[74,] 15 2	[74,] 37 31
[75,] 63 62	[75,] 72 62	[75,] 42 38
[76,] 14 1	[76,] 67 65	[76,] 84 65
[77,] 13 3	[77,] 36 31	[77,] 85 65
[78,] 19 1	[78,] 3 1	[78,] 4 1
[79,] 3 1	[79,] 42 38	[79,] 12 1
[80,] 48 31	[80,] 37 31	[80,] 2 1
[81,] 62 61	[81,] 5 2	[81,] 65 62
[82,] 12 1	[82,] 4 1	[82,] 19 1
[83,] 4 1	[83,] 65 61	[83,] 38 32
[84,] 38 31	[84,] 38 31	[84,] 32 31
[85,] 85 61	[85,] 62 61	[85,] 3 1
[86,] 37 31	[86,] 32 31	[86,] 62 61
[87,] 15 1	[87,] 2 1	[87,] 15 1
[88,] 31 1	[88,] 61 1	[88,] 31 1
[89,] 61 1	[89,] 31 1	[89,] 61 1

2. Data2.csv

simple	complex	average
[,1] [,2]	[,1] [,2]	[,1] [,2]
[1,] 106 9	[1,] 106 9	[1,] 106 9
[2,] 96 79	[2,] 96 79	[2,] 96 79
[3,] 19 4	[3,] 19 4	[3,] 19 4
[4,] 23 3	[4,] 23 3	[4,] 23 3
[5,] 89 67	[5,] 89 67	[5,] 89 67
[6,] 111 6	[6,] 111 6	[6,] 111 6
[7,] 25 2	[7,] 25 2	[7,] 25 2
[8,] 104 7	[8,] 104 7	[8,] 104 7
[9,] 77 70	[9,] 77 70	[9,] 77 70
[10,] 53 48	[10,] 53 48	[10,] 53 48
[11,] 4 3	[11,] 78 72	[11,] 4 3
[12,] 78 72	[12,] 105 14	[12,] 78 72

[13,]	105	14
[14,]	16	5
[15,]	42	40
[16,]	44	36
[17,]	27	15
[18,]	102	71
[19,]	58	54
[20,]	86	76
[21,]	112	52
[22,]	57	54
[23,]	43	37
[24,]	87	80
[25,]	72	67
[26,]	28	13
[27,]	30	14
[28,]	63	62
[29,]	10	1
[30,]	39	37
[31,]	101	54
[32,]	40	33
[33,]	85	62
[34,]	24	6
[35,]	48	47
[36,]	49	38
[37,]	26	12
[38,]	29	5
[39,]	118	54
[40,]	46	33
[41,]	22	9
[42,]	60	33
[43,]	67	62
[44,]	113	33
[45,]	11	8
[46,]	41	38
[47,]	47	35
[48,]	90	81
[49,]	13	8
[50,]	97	83
[51,]	45	31
[52,]	14	1
[53,]	54	31
[54,]	51	33
[55,]	2	1
[56,]	70	62

[13,]	16	5
[14,]	42	40
[15,]	44	36
[16,]	27	15
[17,]	102	71
[18,]	58	54
[19,]	86	76
[20,]	112	52
[21,]	43	37
[22,]	87	80
[23,]	28	13
[24,]	4	3
[25,]	57	54
[26,]	63	62
[27,]	10	1
[28,]	39	37
[29,]	85	62
[30,]	40	33
[31,]	49	38
[32,]	26	12
[33,]	118	101
[34,]	30	14
[35,]	29	5
[36,]	113	60
[37,]	22	9
[38,]	11	8
[39,]	48	47
[40,]	90	81
[41,]	24	6
[42,]	97	83
[43,]	45	31
[44,]	51	46
[45,]	72	67
[46,]	115	110
[47,]	75	65
[48,]	17	2
[49,]	56	36
[50,]	120	103
[51,]	13	8
[52,]	41	38
[53,]	71	66
[54,]	80	68
[55,]	47	32
[56,]	46	34

[13,]	105	14
[14,]	16	5
[15,]	42	40
[16,]	44	36
[17,]	27	15
[18,]	102	71
[19,]	58	54
[20,]	86	76
[21,]	112	52
[22,]	43	37
[23,]	87	80
[24,]	28	13
[25,]	57	54
[26,]	30	14
[27,]	63	62
[28,]	10	1
[29,]	39	37
[30,]	85	62
[31,]	24	6
[32,]	40	33
[33,]	49	38
[34,]	26	12
[35,]	29	5
[36,]	118	101
[37,]	22	9
[38,]	60	33
[39,]	11	8
[40,]	48	47
[41,]	41	38
[42,]	90	81
[43,]	97	83
[44,]	45	31
[45,]	101	54
[46,]	51	46
[47,]	14	2
[48,]	72	67
[49,]	70	67
[50,]	115	110
[51,]	75	65
[52,]	17	2
[53,]	15	3
[54,]	80	68
[55,]	46	33
[56,]	113	33

[57,]	17	1	[57,]	101	54	[57,]	13	1
[58,]	8	1	[58,]	21	18	[58,]	56	36
[59,]	71	66	[59,]	70	69	[59,]	120	103
[60,]	83	82	[60,]	15	3	[60,]	50	36
[61,]	38	31	[61,]	83	79	[61,]	68	62
[62,]	115	110	[62,]	82	74	[62,]	76	62
[63,]	75	65	[63,]	76	62	[63,]	71	66
[64,]	56	36	[64,]	110	1	[64,]	47	32
[65,]	7	1	[65,]	60	33	[65,]	67	62
[66,]	35	32	[66,]	35	32	[66,]	37	32
[67,]	15	1	[67,]	12	2	[67,]	12	2
[68,]	3	1	[68,]	73	64	[68,]	54	34
[69,]	76	62	[69,]	38	31	[69,]	21	18
[70,]	33	31	[70,]	55	52	[70,]	38	31
[71,]	37	32	[71,]	50	36	[71,]	65	61
[72,]	80	68	[72,]	100	93	[72,]	79	61
[73,]	9	1	[73,]	9	8	[73,]	9	8
[74,]	59	31	[74,]	79	61	[74,]	82	74
[75,]	5	1	[75,]	108	33	[75,]	83	61
[76,]	120	103	[76,]	116	107	[76,]	110	1
[77,]	110	1	[77,]	14	1	[77,]	36	34
[78,]	50	36	[78,]	5	3	[78,]	108	33
[79,]	32	31	[79,]	88	81	[79,]	93	5
[80,]	82	62	[80,]	7	2	[80,]	5	3
[81,]	81	66	[81,]	68	62	[81,]	35	32
[82,]	68	62	[82,]	69	67	[82,]	55	52
[83,]	74	62	[83,]	119	117	[83,]	69	66
[84,]	79	62	[84,]	84	61	[84,]	8	1
[85,]	12	1	[85,]	59	31	[85,]	73	64
[86,]	34	31	[86,]	18	2	[86,]	7	2
[87,]	108	31	[87,]	117	99	[87,]	2	1
[88,]	69	62	[88,]	54	34	[88,]	33	31
[89,]	21	1	[89,]	81	66	[89,]	32	31
[90,]	18	1	[90,]	37	32	[90,]	116	107
[91,]	65	61	[91,]	95	94	[91,]	74	62
[92,]	62	61	[92,]	8	1	[92,]	88	81
[93,]	84	61	[93,]	93	3	[93,]	64	62
[94,]	6	1	[94,]	74	62	[94,]	52	34
[95,]	64	61	[95,]	94	65	[95,]	119	117
[96,]	36	31	[96,]	33	31	[96,]	100	3
[97,]	52	31	[97,]	114	103	[97,]	117	99
[98,]	93	1	[98,]	98	20	[98,]	94	61
[99,]	55	31	[99,]	92	91	[99,]	84	61
[100,]	66	61	[100,]	67	62	[100,]	59	31

[101,] 73 61	[101,] 6 1	[101,] 20 3
[102,] 88 61	[102,] 107 103	[102,] 81 66
[103,] 100 1	[103,] 36 34	[103,] 34 31
[104,] 116 107	[104,] 52 34	[104,] 6 1
[105,] 20 1	[105,] 109 99	[105,] 66 62
[106,] 119 117	[106,] 62 61	[106,] 114 103
[107,] 117 99	[107,] 66 64	[107,] 3 1
[108,] 94 61	[108,] 32 31	[108,] 109 99
[109,] 107 31	[109,] 3 2	[109,] 91 18
[110,] 99 31	[110,] 34 31	[110,] 107 103
[111,] 98 31	[111,] 2 1	[111,] 62 61
[112,] 114 31	[112,] 64 61	[112,] 95 61
[113,] 95 1	[113,] 103 20	[113,] 98 31
[114,] 109 31	[114,] 91 65	[114,] 18 1
[115,] 61 1	[115,] 31 20	[115,] 92 61
[116,] 103 31	[116,] 65 1	[116,] 99 31
[117,] 91 1	[117,] 99 20	[117,] 103 31
[118,] 31 1	[118,] 61 1	[118,] 31 1
[119,] 92 1	[119,] 20 1	[119,] 61 1

Every time we put two elements into one cluster, the cluster which has been handled will be represented by the smallest number of the elements in the cluster, then we can print the dendrogram.