Cluster size and VI

2019-1-28

Does k affect VI?

Previously, we applied k-means to divide the variables into 4 groups. Does 4 is a good number to get a good match?

I would like to try how much the number of clusters may affect VI

```
# the known cluster A:
A = rep(1:4,each = 100)

# the other clustering results,
# (i.e. what will happen if we cluster A into k groups)
B = rep(1:2,each = 200)
C = c(1,rep(1:3,each = 133))
D = rep(1:4, each = 100)
E = rep(1:5, each = 80)
f = c(rep(1:6, each = 66),1:4)
G = rep(1:10,each = 40)
H = rep(1:20, each = 20)

vi_random = c()
# Run 1000 times to get a range of VI
vi_random_B = c();vi_random_C = c();vi_random_D = c();vi_random_E = c();
vi_random_F = c();vi_random_G = c();vi_random_H = c()
for(i in 1:1000){
  vi_random_B = c(vi_random,vi(cbind(A,sample(B))))
  vi_random_C = c(vi_random,vi(cbind(A,sample(C))))
  vi_random_D = c(vi_random,vi(cbind(A,sample(D))))
  vi_random_E = c(vi_random,vi(cbind(A,sample(E))))
  vi_random_F = c(vi_random,vi(cbind(A,sample(f))))
  vi_random_G = c(vi_random,vi(cbind(A,sample(G))))
  vi_random_H = c(vi_random,vi(cbind(A,sample(H))))
}

# The results
range_vi = c()
range_vi = rbind(range_vi,range(vi_random_B))
range_vi = rbind(range_vi,range(vi_random_C))
range_vi = rbind(range_vi, range(vi_random_D))
range_vi = rbind(range_vi, range(vi_random_E))
range_vi = rbind(range_vi, range(vi_random_F))
range_vi = rbind(range_vi,range(vi_random_G))
range_vi = rbind(range_vi, range(vi_random_H))

vis = c(vi(cbind(A,B)),vi(cbind(A,C)),vi(cbind(A,D)),vi(cbind(A,E)),
vi(cbind(A,f)),vi(cbind(A,G)),vi(cbind(A,H)))
```

	match	min random	max random
k = 2	0.6931472	2.076840	2.076840

	match	min random	max random
k = 3	0.9252954	2.476447	2.476447
k = 4	0.0000000	2.746865	2.746865
k = 5	0.9502705	2.950780	2.950780
k = 6	1.0101560	3.132244	3.132244
k = 10	1.1935496	3.589462	3.589462
k = 20	1.6094379	4.233322	4.233322

Therefore the VI can be different, in different range. So how could we make them comparable?

Unadjusted VI

Previously, we applied k-means to divide the variables into 4 groups. Does 4 is a good number to get a good match?

To figure it out, I then change k from 2 to 10. Here are some results.

One variable

I then only chose one continuous variable and cluster subjects based on this variable into k groups (k = 2,3,...10). The results:

```
[1] "*****************************************************************"
[1] "k =  2"
[1] "*****************************************************************"
[1] "**** min vi ****"
[1] 0.7249044
[1] "**** the summary table ****"
        cluster 1 cluster 2 Total
Drug            1        71    72
Placebo         0        72    72
Total           1       143   144
[1] "*** p value of the summary table ***"
[1] 1
[1] "*****************************************************************"
[1] "k =  3"
[1] "*****************************************************************"
[1] "**** min vi ****"
[1] 0.8601843
[1] "**** the summary table ****"
        cluster 1 cluster 2 cluster 3 Total
Drug            1         1        73    75
Placebo         2         1        71    74
Total           3         2       144   149
[1] "*** p value of the summary table ***"
[1] 0.8086469
[1] "*****************************************************************"
[1] "k =  4"
[1] "*****************************************************************"
[1] "**** min vi ****"
[1] 1.538022
```

```
[1] "**** the summary table ****"
        cluster 1 cluster 2 cluster 3 cluster 4 Total
Drug            1        31        20        10    62
Placebo         0        26        25         8    59
Total           1        57        45        18   121
[1] "*** p value of the summary table ***"
[1] 0.6004266
[1] "************************************************************"
[1] "k =  5"
[1] "************************************************************"
[1] "**** min vi ****"
[1] 1.644913
[1] "**** the summary table ****"
        cluster 1 cluster 2 cluster 3 cluster 4 cluster 5 Total
Drug           50        19         1         3         3    76
Placebo        47        19         1         5         2    74
Total          97        38         2         8         5   150
[1] "*** p value of the summary table ***"
[1] 0.9393863
[1] "************************************************************"
[1] "k =  6"
[1] "************************************************************"
[1] "**** min vi ****"
[1] 1.860564
[1] "**** the summary table ****"
        cluster 1 cluster 2 cluster 3 cluster 4 cluster 5 cluster 6 Total
Drug            3        16        38         5        13         1    76
Placebo         4        14        39         2        13         2    74
Total           7        30        77         7        26         3   150
[1] "*** p value of the summary table ***"
[1] 0.8937515
[1] "************************************************************"
[1] "k =  7"
[1] "************************************************************"
[1] "**** min vi ****"
[1] 2.028228
[1] "**** the summary table ****"
        cluster 1 cluster 2 cluster 3 cluster 4 cluster 5 cluster 6
Drug           22         1        10        31         0         3
Placebo        25         2         5        35         1         1
Total          47         3        15        66         1         4
        cluster 7 Total
Drug            7    74
Placebo         5    74
Total          12   148
[1] "*** p value of the summary table ***"
[1] 0.6155582
[1] "************************************************************"
[1] "k =  8"
[1] "************************************************************"
[1] "**** min vi ****"
[1] 2.161106
[1] "**** the summary table ****"
        cluster 1 cluster 2 cluster 3 cluster 4 cluster 5 cluster 6
```

```
Drug            5        24         1        10         1         5
Placebo         4        19         1        10         4         6
Total           9        43         2        20         5        11
        cluster 7 cluster 8 Total
Drug            2        13    61
Placebo         5        13    62
Total           7        26   123
[1] "*** p value of the summary table ***"
[1] 0.8186343
[1] "************************************************************"
[1] "k =  9"
[1] "************************************************************"
[1] "**** min vi ****"
[1] 2.381794
[1] "**** the summary table ****"
        cluster 1 cluster 2 cluster 3 cluster 4 cluster 5 cluster 6
Drug            1        16         1         3         6         6
Placebo         0        14         4         3         4        11
Total           1        30         5         6        10        17
        cluster 7 cluster 8 cluster 9 Total
Drug            3        35         5    76
Placebo         1        32         5    74
Total           4        67        10   150
[1] "*** p value of the summary table ***"
[1] 0.7052847
[1] "************************************************************"
[1] "k =  10"
[1] "************************************************************"

Warning: did not converge in 10 iterations

Warning: did not converge in 10 iterations

[1] "**** min vi ****"
[1] 2.325077
[1] "**** the summary table ****"
        cluster 1 cluster 2 cluster 3 cluster 4 cluster 5 cluster 6
Drug           35         0        16         1         5         3
Placebo        32         3        14         0         5         2
Total          67         3        30         1        10         5
        cluster 7 cluster 8 cluster 9 cluster 10 Total
Drug            6         1         6          3    76
Placebo         4         2        11          1    74
Total          10         3        17          4   150
[1] "*** p value of the summary table ***"
[1] 0.6345082
```

Two variables

I then chose two continuous variables and cluster subjects based on this variable into k groups (k = 2,3,. . . 10). The results:

```
[1] "************************************************************"
[1] "k =  2"
[1] "************************************************************"
```

```
[1] "**** min vi ****"
[1] 0.7249044
[1] "**** the summary table ****"
        cluster 1 cluster 2 Total
Drug           32        44    76
Placebo        43        31    74
Total          75        75   150
[1] "*** p value of the summary table ***"
[1] 0.07207462
[1] "************************************************************"
[1] "k =  3"
[1] "************************************************************"
[1] "**** min vi ****"
[1] 0.8567325
[1] "**** the summary table ****"
        cluster 1 cluster 2 cluster 3 Total
Drug           14        19        20    53
Placebo        15        18        23    56
Total          29        37        43   109
[1] "*** p value of the summary table ***"
[1] 0.9696446
[1] "************************************************************"
[1] "k =  4"
[1] "************************************************************"
[1] "**** min vi ****"
[1] 1.45378
[1] "**** the summary table ****"
        cluster 1 cluster 2 cluster 3 cluster 4 Total
Drug            1         4        51        20    76
Placebo         2         4        51        17    74
Total           3         8       102        37   150
[1] "*** p value of the summary table ***"
[1] 0.9370741
[1] "************************************************************"
[1] "k =  5"
[1] "************************************************************"
[1] "**** min vi ****"
[1] 1.609674
[1] "**** the summary table ****"
        cluster 1 cluster 2 cluster 3 cluster 4 cluster 5 Total
Drug           50        19         3         3         1    76
Placebo        47        19         5         2         1    74
Total          97        38         8         5         2   150
[1] "*** p value of the summary table ***"
[1] 0.9393863
[1] "************************************************************"
[1] "k =  6"
[1] "************************************************************"
[1] "**** min vi ****"
[1] 1.757718
[1] "**** the summary table ****"
        cluster 1 cluster 2 cluster 3 cluster 4 cluster 5 cluster 6 Total
Drug           38        18         3        13         1         3    76
Placebo        39        14         2        13         1         5    74
```

```
Total          77        32         5        26         2         8   150
[1] "*** p value of the summary table ***"
[1] 0.9410259
[1] "************************************************************"
[1] "k =  7"
[1] "************************************************************"
[1] "**** min vi ****"
[1] 1.971344
[1] "**** the summary table ****"
        cluster 1 cluster 2 cluster 3 cluster 4 cluster 5 cluster 6
Drug           11         9        13         1        38         3
Placebo        10         7        12         1        39         2
Total          21        16        25         2        77         5
        cluster 7 Total
Drug            1    76
Placebo         3    74
Total           4   150
[1] "*** p value of the summary table ***"
[1] 0.9691799
[1] "************************************************************"
[1] "k =  8"
[1] "************************************************************"
[1] "**** min vi ****"
[1] 2.16204
[1] "**** the summary table ****"
        cluster 1 cluster 2 cluster 3 cluster 4 cluster 5 cluster 6
Drug           10         3         5        16        10        11
Placebo         8         4         6        13        16        11
Total          18         7        11        29        26        22
        cluster 7 cluster 8 Total
Drug            4         5    64
Placebo         9         3    70
Total          13         8   134
[1] "*** p value of the summary table ***"
[1] 0.7539594
[1] "************************************************************"
[1] "k =  9"
[1] "************************************************************"
[1] "**** min vi ****"
[1] 2.263724
[1] "**** the summary table ****"
        cluster 1 cluster 2 cluster 3 cluster 4 cluster 5 cluster 6
Drug            2         7         1        37        13         7
Placebo         2         5         3        38         9        12
Total           4        12         4        75        22        19
        cluster 7 cluster 8 cluster 9 Total
Drug            1         3         5    76
Placebo         1         2         2    74
Total           2         5         7   150
[1] "*** p value of the summary table ***"
[1] 0.7817691
[1] "************************************************************"
[1] "k =  10"
[1] "************************************************************"
```

```
[1] "**** min vi ****"
[1] 2.221511
[1] "**** the summary table ****"
        cluster 1 cluster 2 cluster 3 cluster 4 cluster 5 cluster 6
Drug            6         2         4         5         1         9
Placebo         3         2         3         7         3         9
Total           9         4         7        12         4        18
        cluster 7 cluster 8 cluster 9 cluster 10 Total
Drug            7         5         3          1    43
Placebo         8         6         7          0    48
Total          15        11        10          1    91
[1] "*** p value of the summary table ***"
[1] 0.8781539
```

The summary table

One variable

2	3	4	5	6	7	8	9	10
0.725	0.86	1.538	1.645	1.861	2.028	2.161	2.382	2.325

Two variables

2	3	4	5	6	7	8	9	10
0.725	0.857	1.454	1.61	1.758	1.971	2.162	2.264	2.222

When k $>$ 2, the vi is increasing.

However, they are not comparable, since

Adjusted VI

One variable

I then only chose one continuous variable and cluster subjects based on this variable into k groups (k = 2,3,...10). The results:

```
[1] "************************************************************"
[1] "k =  2"
[1] "************************************************************"
[1] "**** min vi ****"
[1] 0.522908
[1] "**** the summary table ****"
        cluster 1 cluster 2 Total
Drug            1        71    72
Placebo         0        72    72
Total           1       143   144
[1] "*** p value of the summary table ***"
[1] 1
[1] "************************************************************"
```

```
[1] "k =  3"
[1] "************************************************************"
[1] "**** min vi ****"
[1] 0.3914868
[1] "**** the summary table ****"
        cluster 1 cluster 2 cluster 3 Total
Drug            1         1        73    75
Placebo         2         1        71    74
Total           3         2       144   149
[1] "*** p value of the summary table ***"
[1] 0.8086469
[1] "************************************************************"
[1] "k =  4"
[1] "************************************************************"
[1] "**** min vi ****"
[1] 0.5523474
[1] "**** the summary table ****"
        cluster 1 cluster 2 cluster 3 cluster 4 Total
Drug            1        20        51         4    76
Placebo         2        17        51         4    74
Total           3        37       102         8   150
[1] "*** p value of the summary table ***"
[1] 0.9370741
[1] "************************************************************"
[1] "k =  5"
[1] "************************************************************"
[1] "**** min vi ****"
[1] 0.511021
[1] "**** the summary table ****"
        cluster 1 cluster 2 cluster 3 cluster 4 cluster 5 Total
Drug           14        16         4        41         1    76
Placebo         9        16         4        43         2    74
Total          23        32         8        84         3   150
[1] "*** p value of the summary table ***"
[1] 0.8242422
[1] "************************************************************"
[1] "k =  6"
[1] "************************************************************"
[1] "**** min vi ****"
[1] 0.4951194
[1] "**** the summary table ****"
        cluster 1 cluster 2 cluster 3 cluster 4 cluster 5 cluster 6 Total
Drug           47        15         7         1         3         3    76
Placebo        46        15         5         1         5         2    74
Total          93        30        12         2         8         5   150
[1] "*** p value of the summary table ***"
[1] 0.9585771
[1] "************************************************************"
[1] "k =  7"
[1] "************************************************************"
[1] "**** min vi ****"
[1] 0.5071961
[1] "**** the summary table ****"
        cluster 1 cluster 2 cluster 3 cluster 4 cluster 5 cluster 6
```

```
Drug            1        15         7         1         2        47
Placebo         1        15         5         3         2        46
Total           2        30        12         4         4        93
        cluster 7 Total
Drug            3    76
Placebo         2    74
Total           5   150
[1] "*** p value of the summary table ***"
[1] 0.9689525
[1] "************************************************************"
[1] "k =  8"
[1] "************************************************************"

Warning: did not converge in 10 iterations

Warning: did not converge in 10 iterations

[1] "**** min vi ****"
[1] 0.528014
[1] "**** the summary table ****"
        cluster 1 cluster 2 cluster 3 cluster 4 cluster 5 cluster 6
Drug           13        37         1         3         1        10
Placebo         9        38         1         2         3         7
Total          22        75         2         5         4        17
        cluster 7 cluster 8 Total
Drug            7         4    76
Placebo        12         2    74
Total          19         6   150
[1] "*** p value of the summary table ***"
[1] 0.741747
[1] "************************************************************"
[1] "k =  9"
[1] "************************************************************"
[1] "**** min vi ****"
[1] 0.5314324
[1] "**** the summary table ****"
        cluster 1 cluster 2 cluster 3 cluster 4 cluster 5 cluster 6
Drug           35         3         0        14         6         6
Placebo        32         2         3        12        13         5
Total          67         5         3        26        19        11
        cluster 7 cluster 8 cluster 9 Total
Drug            8         1         3    76
Placebo         4         2         1    74
Total          12         3         4   150
[1] "*** p value of the summary table ***"
[1] 0.3890411
[1] "************************************************************"
[1] "k =  10"
[1] "************************************************************"

Warning: did not converge in 10 iterations

[1] "**** min vi ****"
[1] 0.5250155
[1] "**** the summary table ****"
        cluster 1 cluster 2 cluster 3 cluster 4 cluster 5 cluster 6
```

```
Drug            1         3        35         4         3         5
Placebo         4         1        32         7         3         7
Total           5         4        67        11         6        12
        cluster 7 cluster 8 cluster 9 cluster 10 Total
Drug            1        13         6          5    76
Placebo         0        12         3          5    74
Total           1        25         9         10   150
[1] "*** p value of the summary table ***"
[1] 0.7864494
```

Two variables

I then chose two continuous variables and cluster subjects based on this variable into k groups (k = 2,3,. . . 10). The results:

```
[1] "************************************************************"
[1] "k =  2"
[1] "************************************************************"
[1] "**** min vi ****"
[1] 0.522908
[1] "**** the summary table ****"
        cluster 1 cluster 2 Total
Drug           32        44    76
Placebo        43        31    74
Total          75        75   150
[1] "*** p value of the summary table ***"
[1] 0.07207462
[1] "************************************************************"
[1] "k =  3"
[1] "************************************************************"
[1] "**** min vi ****"
[1] 0.3899158
[1] "**** the summary table ****"
        cluster 1 cluster 2 cluster 3 Total
Drug           19        14        20    53
Placebo        18        15        23    56
Total          37        29        43   109
[1] "*** p value of the summary table ***"
[1] 0.9696446
[1] "************************************************************"
[1] "k =  4"
[1] "************************************************************"
[1] "**** min vi ****"
[1] 0.5204815
[1] "**** the summary table ****"
        cluster 1 cluster 2 cluster 3 cluster 4 Total
Drug            4         1        20        51    76
Placebo         4         2        17        51    74
Total           8         3        37       102   150
[1] "*** p value of the summary table ***"
[1] 0.9370741
[1] "************************************************************"
[1] "k =  5"
[1] "************************************************************"
```

```
[1] "**** min vi ****"
[1] 0.4898708
[1] "**** the summary table ****"
        cluster 1 cluster 2 cluster 3 cluster 4 cluster 5 Total
Drug           19         3        50         1         3    76
Placebo        19         2        47         1         5    74
Total          38         5        97         2         8   150
[1] "*** p value of the summary table ***"
[1] 0.9393863
[1] "************************************************************"
[1] "k =  6"
[1] "************************************************************"
[1] "**** min vi ****"
[1] 0.5044645
[1] "**** the summary table ****"
        cluster 1 cluster 2 cluster 3 cluster 4 cluster 5 cluster 6 Total
Drug            1         3         7         3        15        47    76
Placebo         1         2         5         5        15        46    74
Total           2         5        12         8        30        93   150
[1] "*** p value of the summary table ***"
[1] 0.9585771
[1] "************************************************************"
[1] "k =  7"
[1] "************************************************************"
[1] "**** min vi ****"
[1] 0.4987697
[1] "**** the summary table ****"
        cluster 1 cluster 2 cluster 3 cluster 4 cluster 5 cluster 6
Drug            1        41         1        16         4         3
Placebo         3        43         1        16         2         2
Total           4        84         2        32         6         5
        cluster 7 Total
Drug           10    76
Placebo         7    74
Total          17   150
[1] "*** p value of the summary table ***"
[1] 0.9043217
[1] "************************************************************"
[1] "k =  8"
[1] "************************************************************"
[1] "**** min vi ****"
[1] 0.5255245
[1] "**** the summary table ****"
        cluster 1 cluster 2 cluster 3 cluster 4 cluster 5 cluster 6
Drug           11         3         1         7        13         1
Placebo        10         2         3         5        12         1
Total          21         5         4        12        25         2
        cluster 7 cluster 8 Total
Drug            2        38    76
Placebo         2        39    74
Total           4        77   150
[1] "*** p value of the summary table ***"
[1] 0.9852673
[1] "************************************************************"
```

```
[1] "k =  9"
[1] "************************************************************"
[1] "**** min vi ****"
[1] 0.5090848
[1] "**** the summary table ****"
        cluster 1 cluster 2 cluster 3 cluster 4 cluster 5 cluster 6
Drug            1         6         1         5         2        10
Placebo         3         5         0         7         2         9
Total           4        11         1        12         4        19
        cluster 7 cluster 8 cluster 9 Total
Drug            4         7         7    43
Placebo         4        10         8    48
Total           8        17        15    91
[1] "*** p value of the summary table ***"
[1] 0.980098
[1] "************************************************************"
[1] "k =  10"
[1] "************************************************************"
[1] "**** min vi ****"
[1] 0.5224525
[1] "**** the summary table ****"
        cluster 1 cluster 2 cluster 3 cluster 4 cluster 5 cluster 6
Drug           15        13         5         5         4         5
Placebo        11        10         6        11         3         6
Total          26        23        11        16         7        11
        cluster 7 cluster 8 cluster 9 cluster 10 Total
Drug            4        19         3          2    75
Placebo         0        21         4          2    74
Total           4        40         7          4   149
[1] "*** p value of the summary table ***"
[1] 0.5708574
```

The summary table

One variable

2	3	4	5	6	7	8	9	10
0.523	0.391	0.552	0.511	0.495	0.507	0.528	0.531	0.525

Two variables

2	3	4	5	6	7	8	9	10
0.523	0.39	0.52	0.49	0.504	0.499	0.526	0.509	0.522

When VI can be the max?

Overall, it seems that, when the number of cluster matches, the VI can achive the max value.