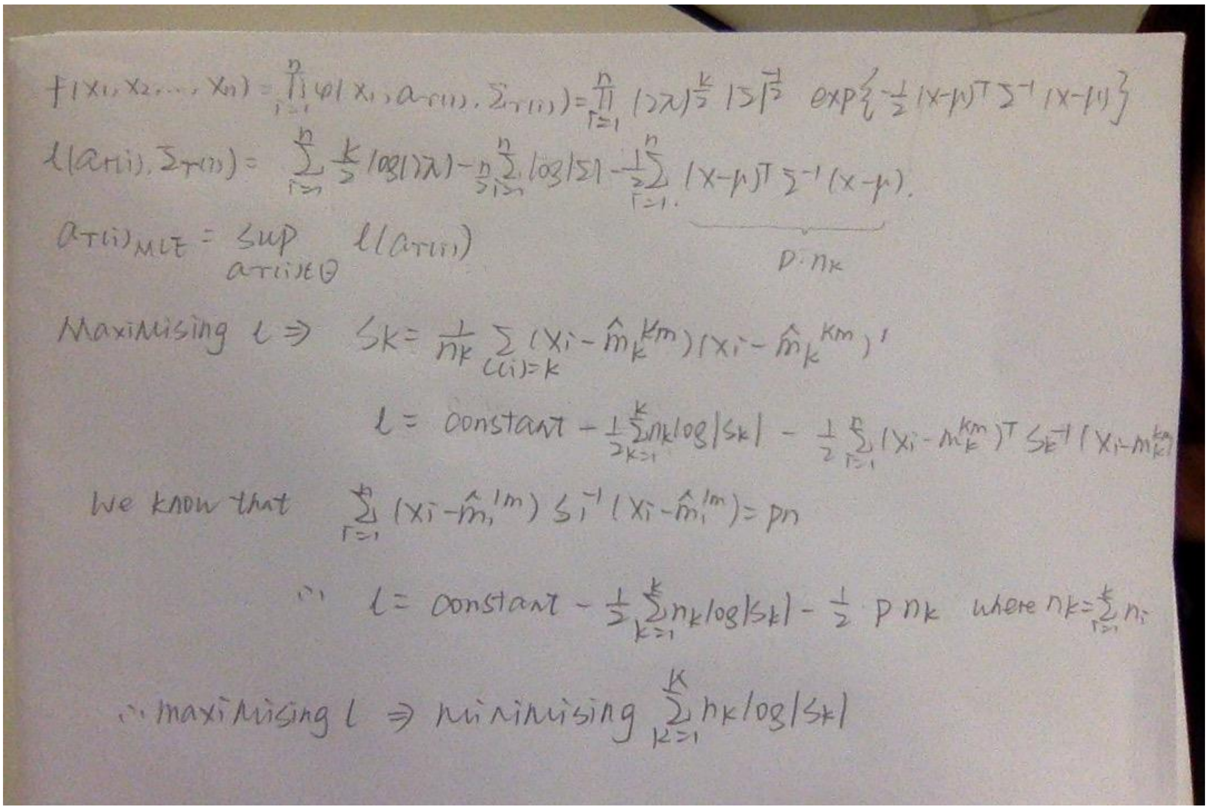
Name: Yina Lin

STAT 3019 Excersice 2

Question 1:



$f(x_1,x_2,\dots,x_n)=\prod_{i=1}^{n}\varphi(x_i,a_{r(i)},\Sigma_{r(i)})=\prod_{i=1}^{n}(2\pi)^{\frac{k}{2}}|\Sigma|^{-\frac{1}{2}}\exp\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\}$

$l(a_{r(i)},\Sigma_{r(i)})=\sum_{i=1}^{n}\frac{k}{2}\log(2\pi)-\frac{n}{2}\sum_{i=1}^{n}\log|\Sigma|-\frac{1}{2}\sum_{i=1}^{n}\underbrace{(x-\mu)^T\Sigma^{-1}(x-\mu)}_{p\cdot n_k}.$

$a_{r(i)MLE}=\sup_{a_{r(i)}\in\Theta}l(a_{r(i)})$

Maximising $l \Rightarrow S_k=\frac{1}{n_k}\sum_{c(i)=k}(x_i-\hat{m}_k^{km})(x_i-\hat{m}_k^{km})'$

$l=\text{constant}-\frac{1}{2}\sum_{k=1}^{k}n_k\log|S_k|-\frac{1}{2}\sum_{i=1}^{n}(x_i-m_k^{km})^T S_k^{-1}(x_i-m_k^{km}$

We know that $\sum_{i=1}^{n}(x_i-\hat{m}_i^{lm})S_i^{-1}(x_i-\hat{m}_i^{lm})=pn$

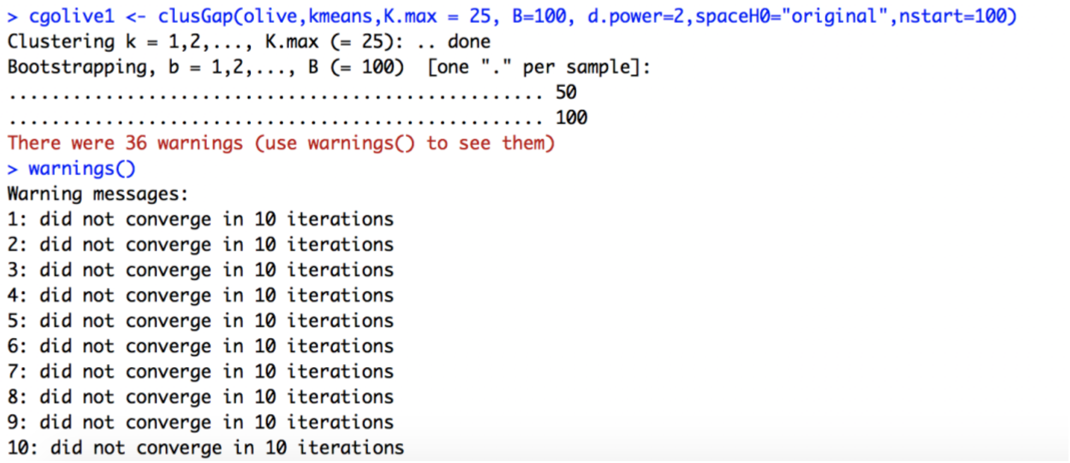
$\therefore\ l=\text{constant}-\frac{1}{2}\sum_{k=1}^{k}n_k\log|S_k|-\frac{1}{2}pn_k$ where $n_k=\sum_{i=1}^{k}n_i$

$\therefore$ maximising $l \Rightarrow$ minimising $\sum_{k=1}^{K}n_k\log|S_k|$

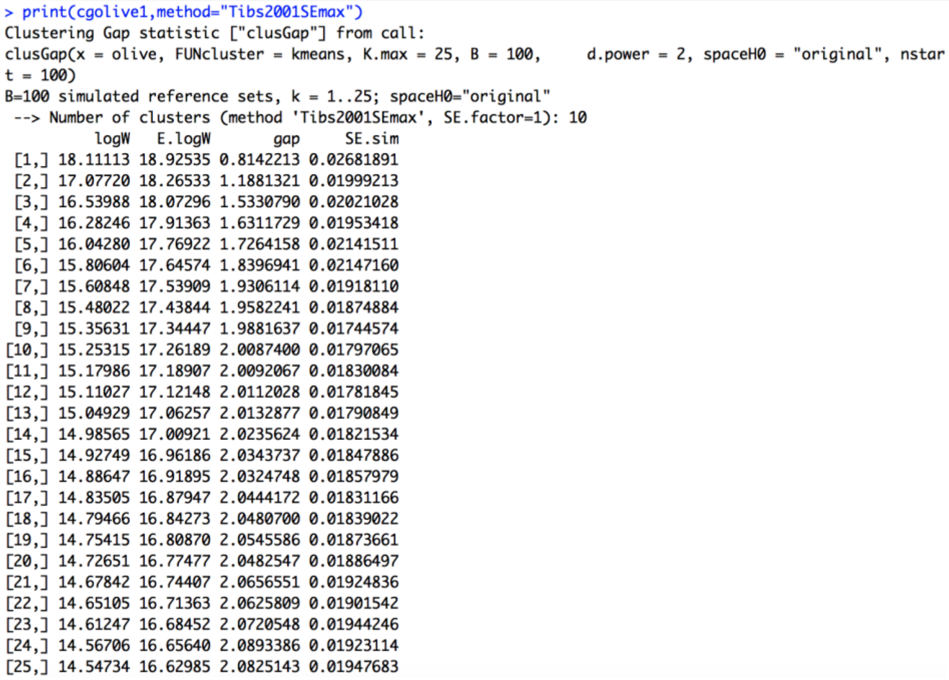
Question 2:

```
>R
set.seed(123456)
cgolive1 <- clusGap(olive,kmeans,K.max = 25, B=100,
d.power=2,spaceH0="original",nstart=100)
print(cg1,method="Tibs2001SEmax")
```

Name: Yina Lin

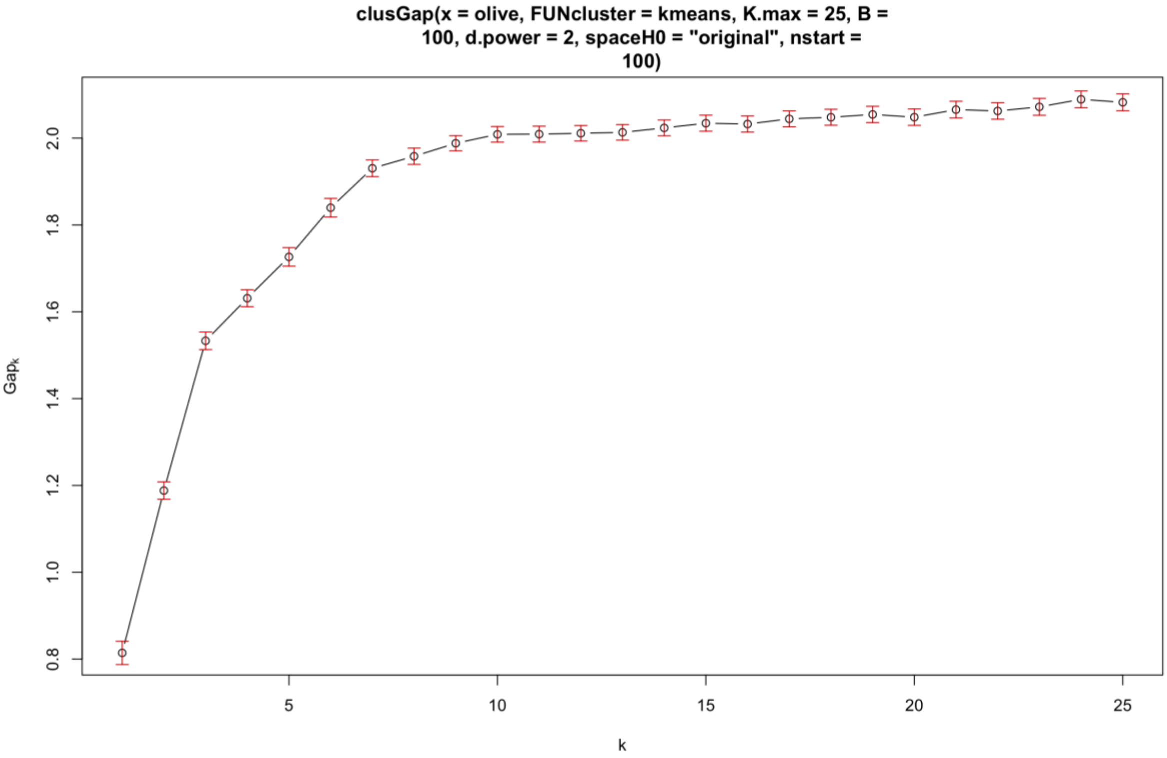


```
> cgolive1 <- clusGap(olive,kmeans,K.max = 25, B=100, d.power=2,spaceH0="original",nstart=100)
Clustering k = 1,2,..., K.max (= 25): .. done
Bootstrapping, b = 1,2,..., B (= 100)  [one "." per sample]:
.................................................. 50
.................................................. 100
There were 36 warnings (use warnings() to see them)
> warnings()
Warning messages:
1: did not converge in 10 iterations
2: did not converge in 10 iterations
3: did not converge in 10 iterations
4: did not converge in 10 iterations
5: did not converge in 10 iterations
6: did not converge in 10 iterations
7: did not converge in 10 iterations
8: did not converge in 10 iterations
9: did not converge in 10 iterations
10: did not converge in 10 iterations
```

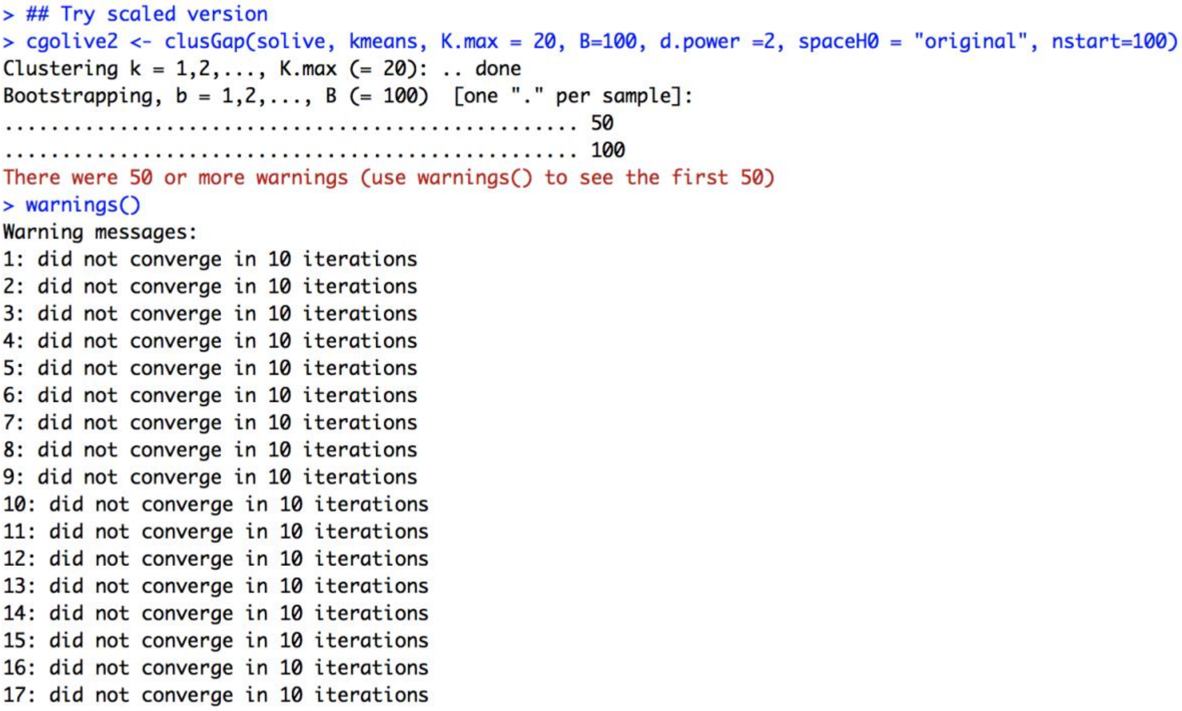


```
> print(cgolive1,method="Tibs2001SEmax")
Clustering Gap statistic ["clusGap"] from call:
clusGap(x = olive, FUNcluster = kmeans, K.max = 25, B = 100,     d.power = 2, spaceH0 = "original", nstar
t = 100)
B=100 simulated reference sets, k = 1..25; spaceH0="original"
 --> Number of clusters (method 'Tibs2001SEmax', SE.factor=1): 10
          logW   E.logW       gap     SE.sim
 [1,] 18.11113 18.92535 0.8142213 0.02681891
 [2,] 17.07720 18.26533 1.1881321 0.01999213
 [3,] 16.53988 18.07296 1.5330790 0.02021028
 [4,] 16.28246 17.91363 1.6311729 0.01953418
 [5,] 16.04280 17.76922 1.7264158 0.02141511
 [6,] 15.80604 17.64574 1.8396941 0.02147160
 [7,] 15.60848 17.53909 1.9306114 0.01918110
 [8,] 15.48022 17.43844 1.9582241 0.01874884
 [9,] 15.35631 17.34447 1.9881637 0.01744574
[10,] 15.25315 17.26189 2.0087400 0.01797065
[11,] 15.17986 17.18907 2.0092067 0.01830084
[12,] 15.11027 17.12148 2.0112028 0.01781845
[13,] 15.04929 17.06257 2.0132877 0.01790849
[14,] 14.98565 17.00921 2.0235624 0.01821534
[15,] 14.92749 16.96186 2.0343737 0.01847886
[16,] 14.88647 16.91895 2.0324748 0.01857979
[17,] 14.83505 16.87947 2.0444172 0.01831166
[18,] 14.79466 16.84273 2.0480700 0.01839022
[19,] 14.75415 16.80870 2.0545586 0.01873661
[20,] 14.72651 16.77477 2.0482547 0.01886497
[21,] 14.67842 16.74407 2.0656551 0.01924836
[22,] 14.65105 16.71363 2.0625809 0.01901542
[23,] 14.61247 16.68452 2.0720548 0.01944246
[24,] 14.56706 16.65640 2.0893386 0.01923114
[25,] 14.54734 16.62985 2.0825143 0.01947683
```

Name: Yina Lin

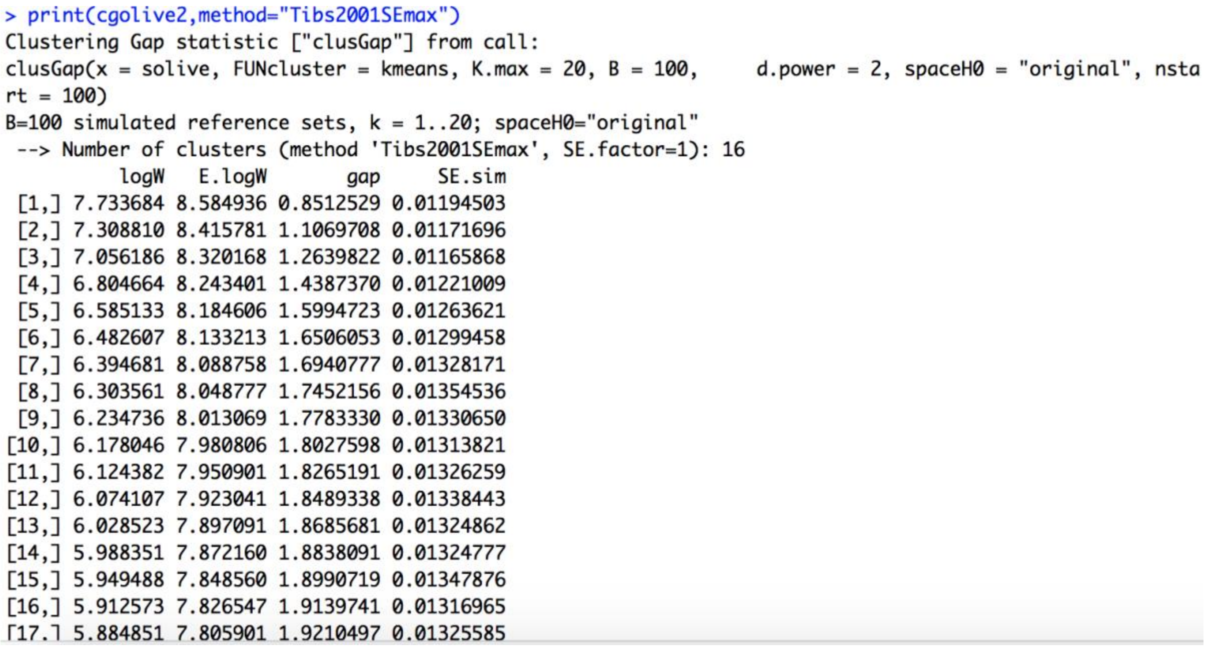


Try scaled version:

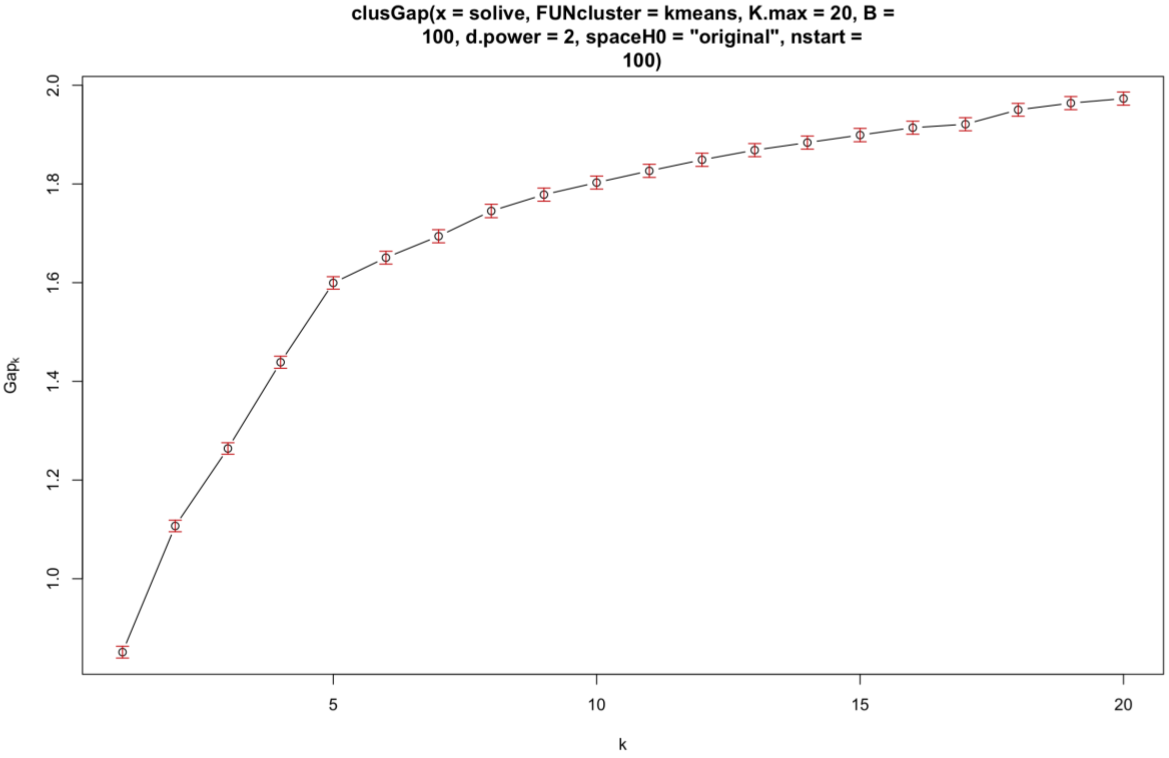


```
> ## Try scaled version
> cgolive2 <- clusGap(solive, kmeans, K.max = 20, B=100, d.power =2, spaceH0 = "original", nstart=100)
Clustering k = 1,2,..., K.max (= 20): .. done
Bootstrapping, b = 1,2,..., B (= 100)  [one "." per sample]:
.................................................. 50
.................................................. 100
There were 50 or more warnings (use warnings() to see the first 50)
> warnings()
Warning messages:
1: did not converge in 10 iterations
2: did not converge in 10 iterations
3: did not converge in 10 iterations
4: did not converge in 10 iterations
5: did not converge in 10 iterations
6: did not converge in 10 iterations
7: did not converge in 10 iterations
8: did not converge in 10 iterations
9: did not converge in 10 iterations
10: did not converge in 10 iterations
11: did not converge in 10 iterations
12: did not converge in 10 iterations
13: did not converge in 10 iterations
14: did not converge in 10 iterations
15: did not converge in 10 iterations
16: did not converge in 10 iterations
17: did not converge in 10 iterations
```

Name: Yina Lin

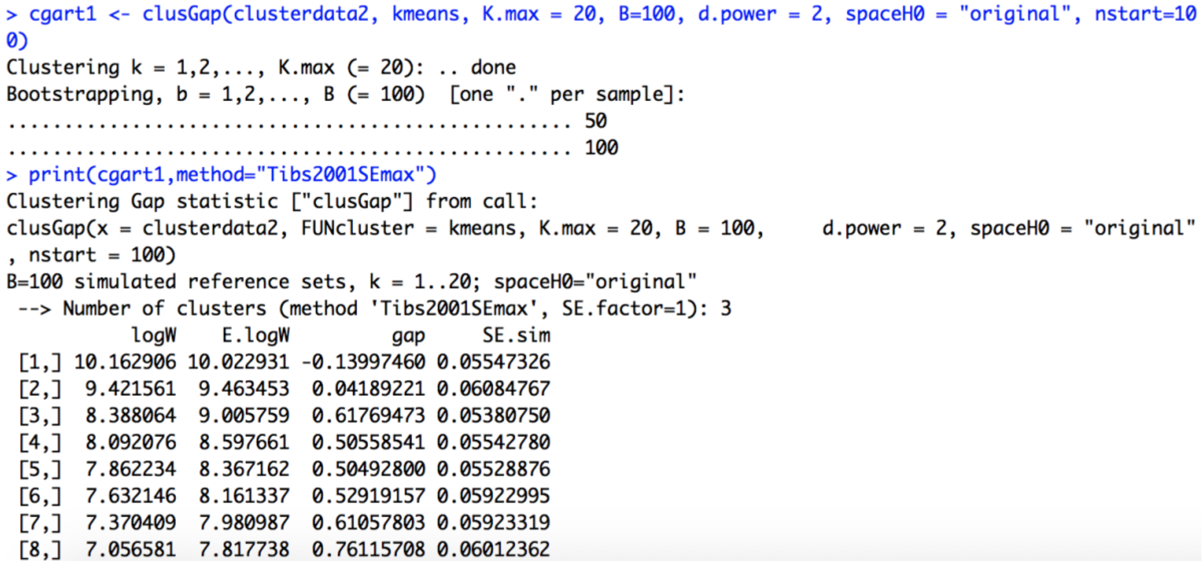


```
> print(cgolive2,method="Tibs2001SEmax")
Clustering Gap statistic ["clusGap"] from call:
clusGap(x = solive, FUNcluster = kmeans, K.max = 20, B = 100,     d.power = 2, spaceH0 = "original", nsta
rt = 100)
B=100 simulated reference sets, k = 1..20; spaceH0="original"
 --> Number of clusters (method 'Tibs2001SEmax', SE.factor=1): 16
          logW   E.logW       gap     SE.sim
 [1,] 7.733684 8.584936 0.8512529 0.01194503
 [2,] 7.308810 8.415781 1.1069708 0.01171696
 [3,] 7.056186 8.320168 1.2639822 0.01165868
 [4,] 6.804664 8.243401 1.4387370 0.01221009
 [5,] 6.585133 8.184606 1.5994723 0.01263621
 [6,] 6.482607 8.133213 1.6506053 0.01299458
 [7,] 6.394681 8.088758 1.6940777 0.01328171
 [8,] 6.303561 8.048777 1.7452156 0.01354536
 [9,] 6.234736 8.013069 1.7783330 0.01330650
[10,] 6.178046 7.980806 1.8027598 0.01313821
[11,] 6.124382 7.950901 1.8265191 0.01326259
[12,] 6.074107 7.923041 1.8489338 0.01338443
[13,] 6.028523 7.897091 1.8685681 0.01324862
[14,] 5.988351 7.872160 1.8838091 0.01324777
[15,] 5.949488 7.848560 1.8990719 0.01347876
[16,] 5.912573 7.826547 1.9139741 0.01316965
[17,] 5.884851 7.805901 1.9210497 0.01325585
```

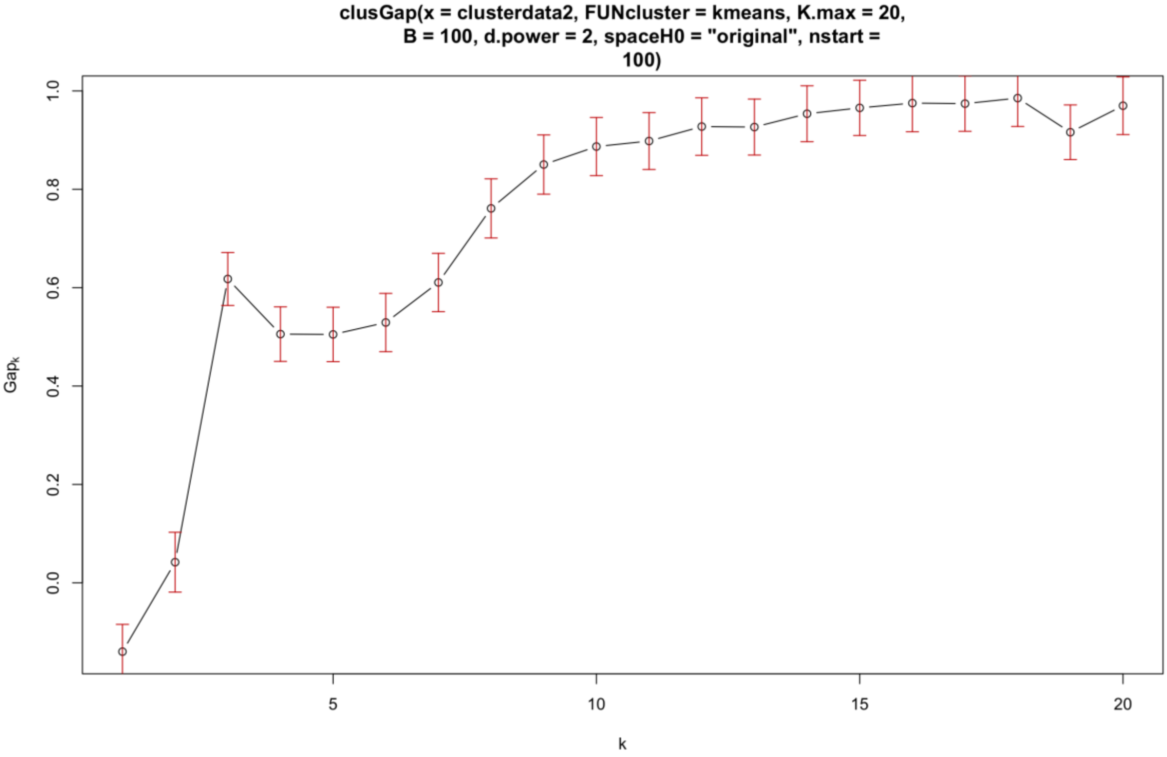


Artificial Dataset2

Name: Yina Lin



```
> cgart1 <- clusGap(clusterdata2, kmeans, K.max = 20, B=100, d.power = 2, spaceH0 = "original", nstart=10
0)
Clustering k = 1,2,..., K.max (= 20): .. done
Bootstrapping, b = 1,2,..., B (= 100)  [one "." per sample]:
.................................................. 50
.................................................. 100
> print(cgart1,method="Tibs2001SEmax")
Clustering Gap statistic ["clusGap"] from call:
clusGap(x = clusterdata2, FUNcluster = kmeans, K.max = 20, B = 100,     d.power = 2, spaceH0 = "original"
, nstart = 100)
B=100 simulated reference sets, k = 1..20; spaceH0="original"
 --> Number of clusters (method 'Tibs2001SEmax', SE.factor=1): 3
          logW    E.logW         gap     SE.sim
 [1,] 10.162906 10.022931 -0.13997460 0.05547326
 [2,]  9.421561  9.463453  0.04189221 0.06084767
 [3,]  8.388064  9.005759  0.61769473 0.05380750
 [4,]  8.092076  8.597661  0.50558541 0.05542780
 [5,]  7.862234  8.367162  0.50492800 0.05528876
 [6,]  7.632146  8.161337  0.52919157 0.05922995
 [7,]  7.370409  7.980987  0.61057803 0.05923319
 [8,]  7.056581  7.817738  0.76115708 0.06012362
```



2nd time:

Clustering Gap statistic ["clusGap"] from call:

clusGap(x = clusterdata2, FUNcluster = kmeans, K.max = 20, B = 100,   d.power = 2, spaceH0 = "original", nstart = 100)

B=100 simulated reference sets, k = 1..20; spaceH0="original"

--> Number of clusters (method 'Tibs2001SEmax', SE.factor=1): 3

```
         logW    E.logW       gap     SE.sim
[1,] 10.162906 10.029403 -0.1335028 0.05459227
[2,]  9.421561  9.469605  0.0480442 0.05835591
[3,]  8.388064  9.009728  0.6216637 0.05899336
[4,]  8.092076  8.607358  0.5152824 0.05104977
[5,]  7.862234  8.374322  0.5120884 0.05589792
[6,]  7.632146  8.170778  0.5386323 0.05540255
[7,]  7.370409  7.986016  0.6156064 0.05452030
```

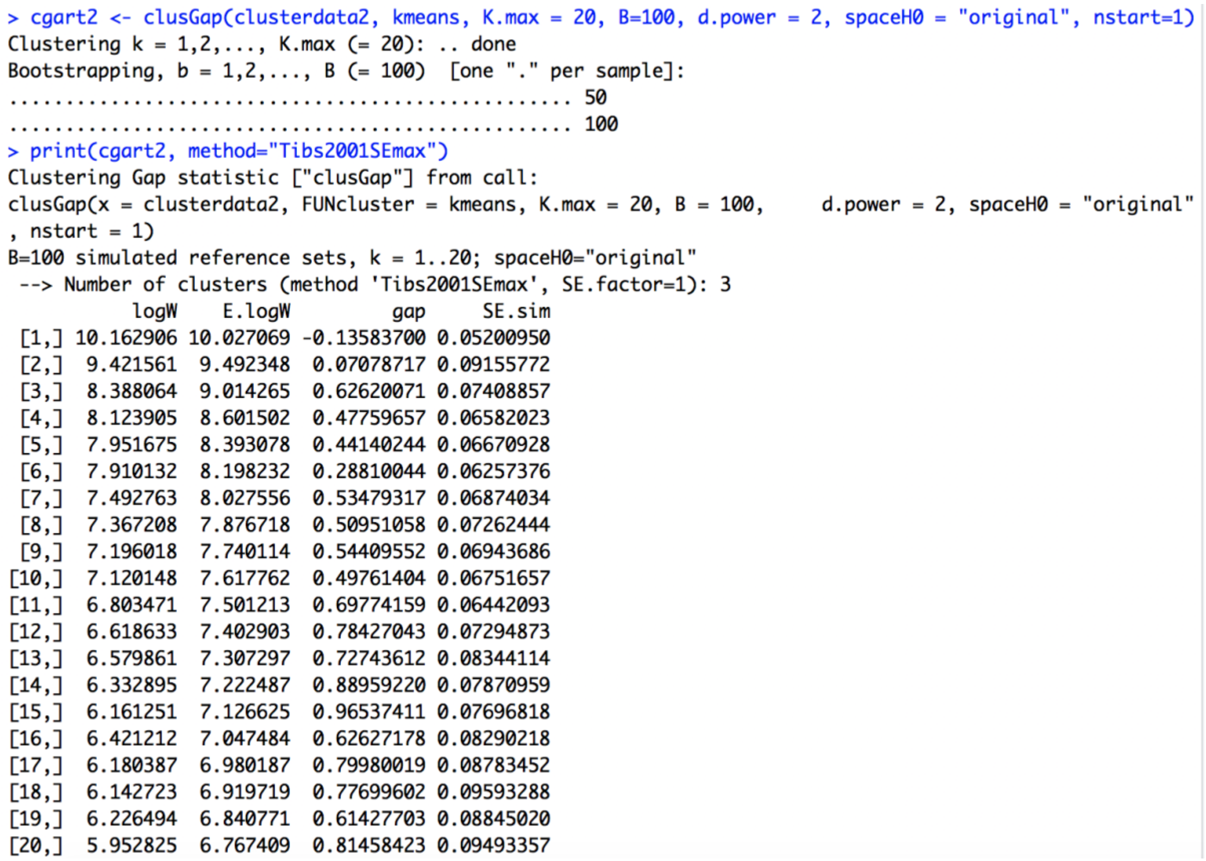
Name: Yina Lin

```
 [8,]  7.056581  7.822098  0.7655178 0.05122344
 [9,]  6.819451  7.674705  0.8552532 0.05177619
[10,]  6.653562  7.544946  0.8913842 0.05169753
[11,]  6.526397  7.425600  0.8992030 0.05481559
[12,]  6.387970  7.314653  0.9266830 0.05500285
[13,]  6.257471  7.213332  0.9558614 0.05499722
[14,]  6.172711  7.117139  0.9444282 0.05268865
[15,]  6.059074  7.028204  0.9691297 0.05300277
[16,]  5.995895  6.946897  0.9510025 0.05035853
[17,]  5.924015  6.866469  0.9424542 0.04896339
[18,]  5.873572  6.793493  0.9199207 0.04896972
[19,]  5.738782  6.718450  0.9796680 0.04910523
[20,]  5.717040  6.647706  0.9306669 0.04936856
```

Same

3rd time:
same

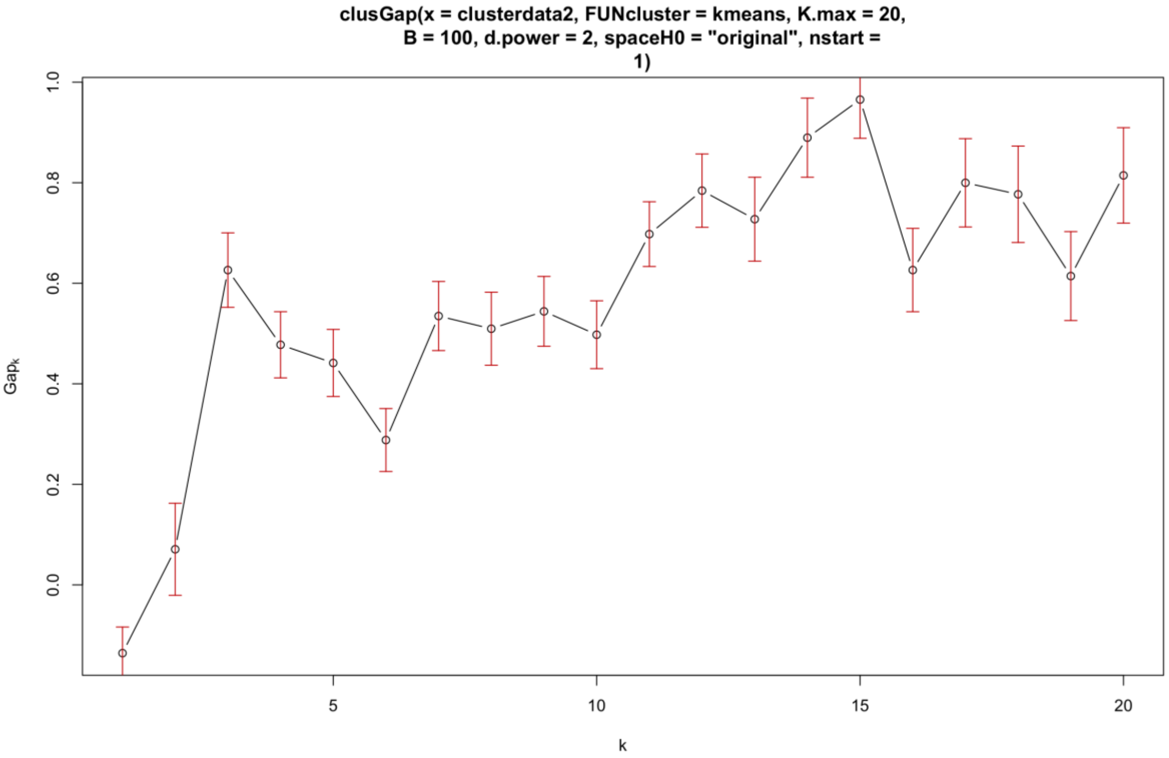
The results are the same, maybe because the nstart that I chose is large enough.



```
> cgart2 <- clusGap(clusterdata2, kmeans, K.max = 20, B=100, d.power = 2, spaceH0 = "original", nstart=1)
Clustering k = 1,2,..., K.max (= 20): .. done
Bootstrapping, b = 1,2,..., B (= 100)  [one "." per sample]:
.................................................. 50
.................................................. 100
> print(cgart2, method="Tibs2001SEmax")
Clustering Gap statistic ["clusGap"] from call:
clusGap(x = clusterdata2, FUNcluster = kmeans, K.max = 20, B = 100,     d.power = 2, spaceH0 = "original"
, nstart = 1)
B=100 simulated reference sets, k = 1..20; spaceH0="original"
 --> Number of clusters (method 'Tibs2001SEmax', SE.factor=1): 3
          logW    E.logW         gap     SE.sim
 [1,] 10.162906 10.027069 -0.13583700 0.05200950
 [2,]  9.421561  9.492348  0.07078717 0.09155772
 [3,]  8.388064  9.014265  0.62620071 0.07408857
 [4,]  8.123905  8.601502  0.47759657 0.06582023
 [5,]  7.951675  8.393078  0.44140244 0.06670928
 [6,]  7.910132  8.198232  0.28810044 0.06257376
 [7,]  7.492763  8.027556  0.53479317 0.06874034
 [8,]  7.367208  7.876718  0.50951058 0.07262444
 [9,]  7.196018  7.740114  0.54409552 0.06943686
[10,]  7.120148  7.617762  0.49761404 0.06751657
[11,]  6.803471  7.501213  0.69774159 0.06442093
[12,]  6.618633  7.402903  0.78427043 0.07294873
[13,]  6.579861  7.307297  0.72743612 0.08344114
[14,]  6.332895  7.222487  0.88959220 0.07870959
[15,]  6.161251  7.126625  0.96537411 0.07696818
[16,]  6.421212  7.047484  0.62627178 0.08290218
[17,]  6.180387  6.980187  0.79980019 0.08783452
[18,]  6.142723  6.919719  0.77699602 0.09593288
[19,]  6.226494  6.840771  0.61427703 0.08845020
[20,]  5.952825  6.767409  0.81458423 0.09493357
```

Hah! Some differences appears when I changed the nstart to 1.

Name: Yina Lin



Yea quite lot differences.

Conclusion: set a high nstart to gain a relatively stable result.

Question 3:

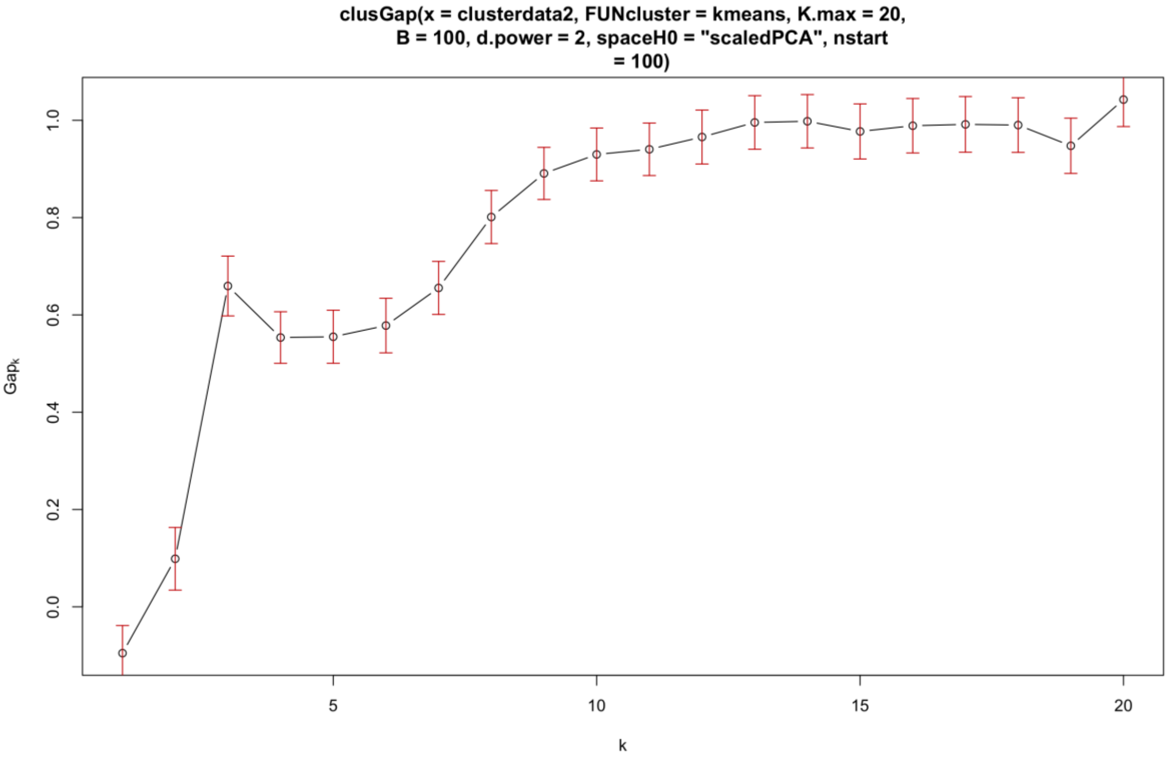
Let's simulate an experimental dataset with a known cluster, say 5, from 5 different distributions. And then tried the different methods to see which one performs better.

Well maybe just used the artificial dataset 2 for a try!

- R

```
set.seed(1234567)
cgart3 <- clusGap(clusterdata2, kmeans, K.max = 20, B=100, d.power = 2,spaceH0 =
"scaledPCA",nstart=100)
print(cgart3,method="Tibs2001SEmax")
plot(cgart3)
```

Name: Yina Lin



There ain’t that much difference for orginal and scaledPCA.

```
> print(cgart1, method= "firstSEmax")
Clustering Gap statistic ["clusGap"] from call:
clusGap(x = clusterdata2, FUNcluster = kmeans, K.max = 20, B = 100,     d.power = 2, spaceH0 = "original",
nstart = 100)
B=100 simulated reference sets, k = 1..20; spaceH0="original"
 --> Number of clusters (method 'firstSEmax', SE.factor=1): 3
          logW    E.logW        gap     SE.sim
 [1,] 10.162906 10.030573 -0.13233240 0.05178483
 [2,]  9.421561  9.472113  0.05055164 0.06070778
 [3,]  8.388064  9.011571  0.62350699 0.04647102
 [4,]  8.092076  8.596701  0.50462502 0.05406854
 [5,]  7.862234  8.369249  0.50701579 0.05252120
 [6,]  7.632146  8.163635  0.53148963 0.05172188
 [7,]  7.370409  7.982243  0.61183364 0.05057345
 [8,]  7.056581  7.814050  0.75746882 0.04523709
 [9,]  6.819451  7.666732  0.84728009 0.04328101
[10,]  6.653562  7.536085  0.88252315 0.04420563
[11,]  6.526397  7.415392  0.88899506 0.04650629
[12,]  6.387970  7.305939  0.91796935 0.04750822
[13,]  6.291425  7.202226  0.91080104 0.04946357
[14,]  6.200447  7.108344  0.90789681 0.04965334
[15,]  6.097031  7.019166  0.92213499 0.05033007
[16,]  5.998996  6.933476  0.93448022 0.05316976
[17,]  5.942038  6.852122  0.91008442 0.05410203
[18,]  5.802220  6.777149  0.97492881 0.05687920
[19,]  5.800740  6.706896  0.90615646 0.05631342
[20,]  5.732909  6.637159  0.90425018 0.05933496
```

There are different results but the numbers of clusters found are the same.

Name: Yina Lin

Question 4: