Statistics 5401
32. K-Means Clustering

Gary W. Oehlert
School of Statistics
313B Ford Hall
612-625-1557
gary@stat.umn.edu

K-means is a nonhierarchical clustering method. You tell it how many clusters you want, and it tries to find the "best" clustering.

K-means is a combinatorially difficult problem, and most algorithms only find approximately optimal clusters.

Here is the idea. There is an "error sum of squares" type criterion we are trying to minimize.

Do a one-way MANOVA of all the data with the clusters as the groups and the the within groups (error) sum of squares and cross products matrix $\mathbf{W}$. Our criterion is the trace of $\mathbf{W}$, which we want to make as small as possible.

What the algorithms do is take an initial clustering into k groups, and then transfer cases between groups to make the criterion smaller. When no futher transfers lower the criterion, we have our final clustering.

In fact, the criterion can have *many* local minima, so it is often best to try several different initial clusterings.

There are several variations on the algorithm, depending on how the transfers are done. The book gives the simplest possible version.

1. Compute the means of each cluster.
2. Cycle through all cases. Allocate case i to cluster j if the mean of cluster j is the closest cluster mean. Recompute means after each transfer.
3. Repeat until each point is in the cluster that has the closest mean.

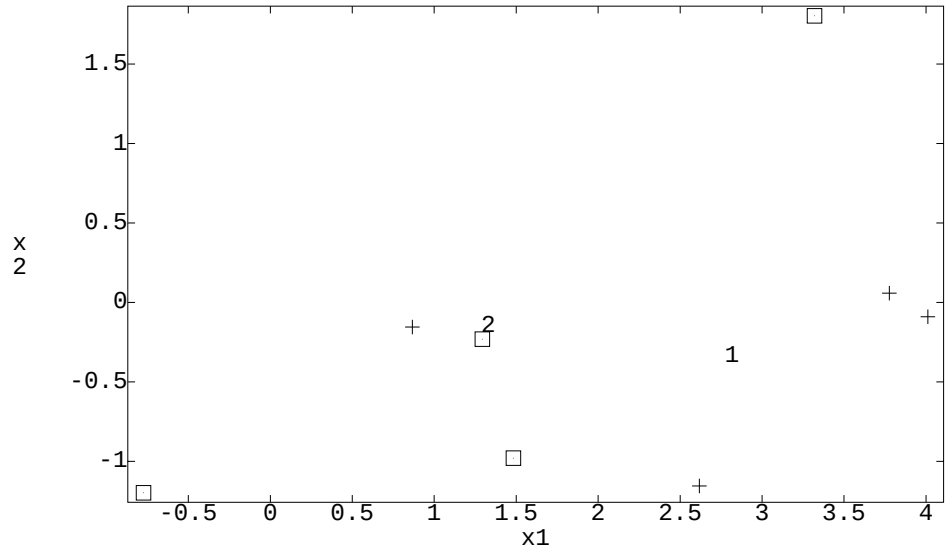
In fact, it is easy to do better than the book's algorithm.

1. Cycle through all cases.
2. For each case, compute the cluster means with that case deleted. Allocate the case to the cluster with the closest mean.
3. Repeat until each point is in the cluster that has the closest mean.

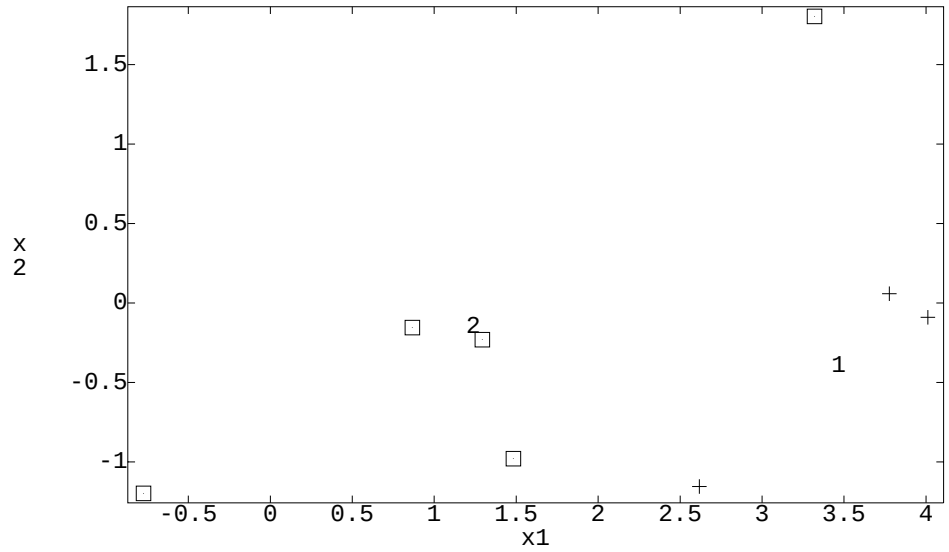
This works noticably better.

```
Cmd> setseeds(1341343222 , 1032315086)

Cmd> x1 <- rnorm(8); x2 <- rnorm(8)

Cmd> x1[run(5,8)] <- x1[run(5,8)]+3

Cmd> clusters <- 1+(grade(runi(8)) < 4.5)

Cmd> clusters
(1)     2     1     2     2     2
(6)     1     1     1

Cmd> mns <- tabs(X,clusters,mean:T)
```

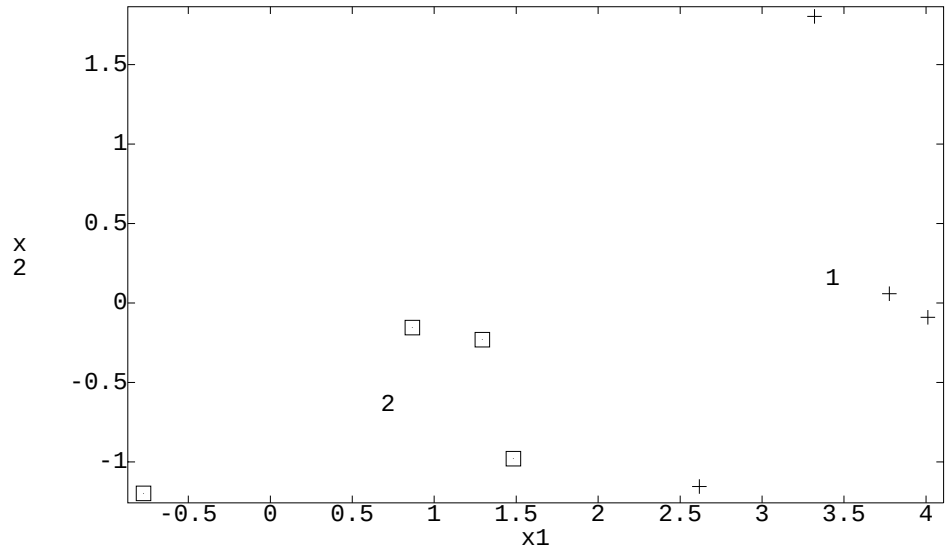
```
Cmd> chplot(x1,x2,\
symbols:vector("\002","\003")[clusters],\
xaxis:F,yaxis:F);plot(mns[,1],mns[,2],\
symbols:vector(1,2),add:T)
```



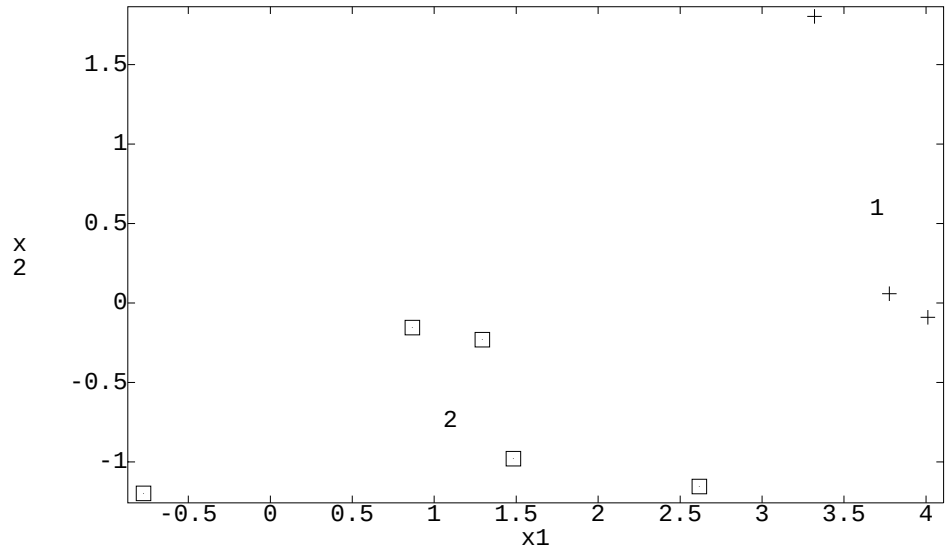

```
Cmd> mnsb <- tabs(X[-1,],clusters[-1],mean:T)

Cmd> sum((mnsb'-X[1,]')^2)
(1,1)        13.64       9.8182

Cmd> mnsb <- tabs(X[-2,],clusters[-2],mean:T)

Cmd> sum((mnsb'-X[2,]')^2)
(1,1)       6.8296      0.21587

Cmd> clusters[2] <- 2

Cmd> mns <- tabs(X,clusters,mean:T)

Cmd> chplot(x1,x2,\
symbols:vector("\002","\003")[clusters],\
xaxis:F,yaxis:F);plot(mns[,1],mns[,2],\
symbols:vector(1,2),add:T)
```




```
Cmd> mnsb <- tabs(X[-3,],clusters[-3],mean:T)

Cmd> sum((mnsb'-X[3,]')^2)
(1,1)       4.2812       1.1632

Cmd> mnsb <- tabs(X[-4,],clusters[-4],mean:T)

Cmd> sum((mnsb'-X[4,]')^2)
(1,1)       4.7562     0.014702

Cmd> mnsb <- tabs(X[-5,],clusters[-5],mean:T)

Cmd> sum((mnsb'-X[5,]')^2)
(1,1)       4.8573       12.743

Cmd> clusters[5] <- 1

Cmd> mns <- tabs(X,clusters,mean:T)

Cmd> chplot(x1,x2,\
symbols:vector("\002","\003")[clusters],\
xaxis:F,yaxis:F);plot(mns[,1],mns[,2],\
symbols:vector(1,2),add:T)
```

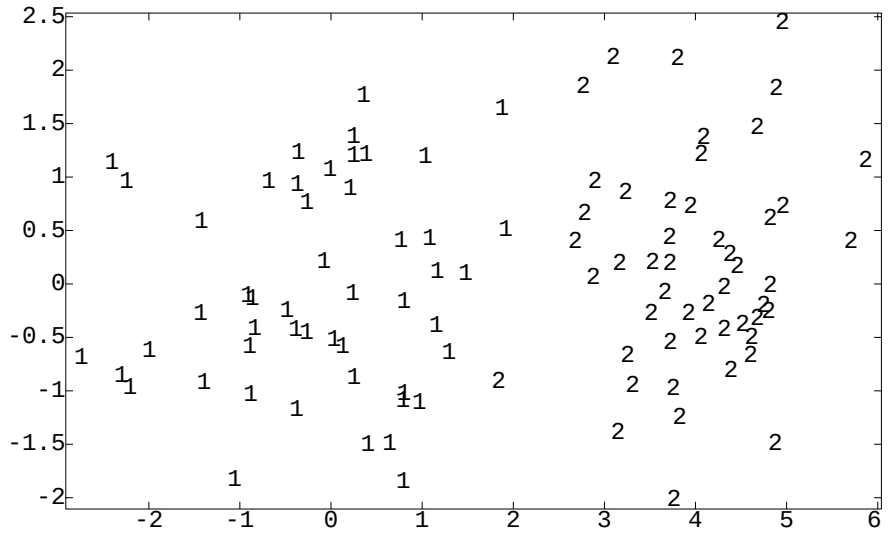



```
Cmd> mnsb <- tabs(X[-6,],clusters[-6],mean:T)

Cmd> sum((mnsb'-X[6,]')^2)
(1,1)         4.22       3.8762

Cmd> clusters[6] <- 2

Cmd> mns <- tabs(X,clusters,mean:T)

Cmd> chplot(x1,x2,\
symbols:vector("\002","\003")[clusters],\
xaxis:F,yaxis:F);plot(mns[,1],mns[,2],\
symbols:vector(1,2),add:T)
```

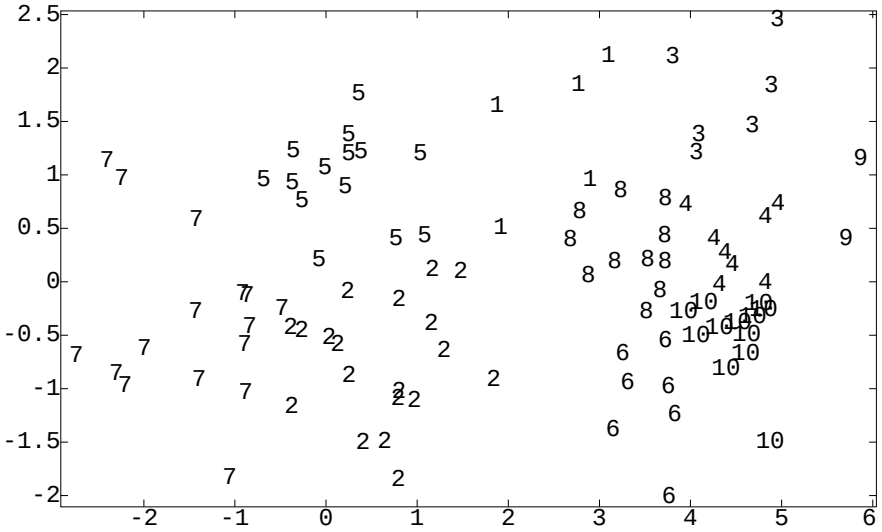



```
Cmd> mnsb <- tabs(X[-7,],clusters[-7],mean:T)

Cmd> sum((mnsb'-X[7,]')^2)
(1,1)       1.2578       8.9145

Cmd> mnsb <- tabs(X[-8,],clusters[-8],mean:T)

Cmd> sum((mnsb'-X[8,]')^2)
(1,1)      0.64919       7.8173

Cmd> clusters
(1)     2     2     2     2     1
(6)     2     1     1

Cmd> for(i,run(8)) {
mnsb <- tabs(X[-i,],clusters[-i],mean:T)
sum((mnsb'-X[i,]')^2)
}
(1,1)       23.228       5.7917
(1,1)       8.5979      0.62546
(1,1)       7.3867      0.31915
(1,1)       6.4756      0.47015
(1,1)       3.6399       11.423
(1,1)         4.22       3.8762
(1,1)       1.2578       8.9145
(1,1)      0.64919       7.8173

Cmd> X <- matrix(rnorm(200),100)
```

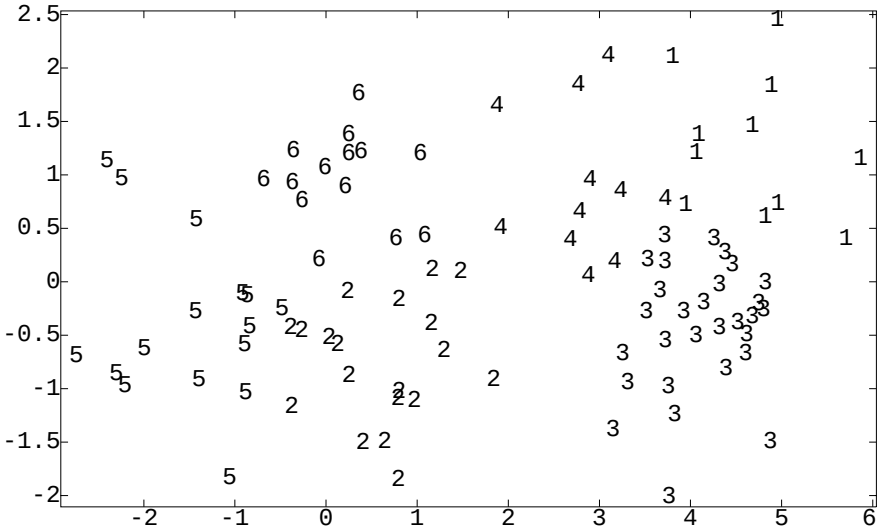
```
Cmd> X[run(51,100),1] <- X[run(51,100),1] + 4

Cmd> tclus <- rep(run(2),rep(50,2))

Cmd> chplot(X[,1],X[,2],tclus,xaxis:F,yaxis:F)
```



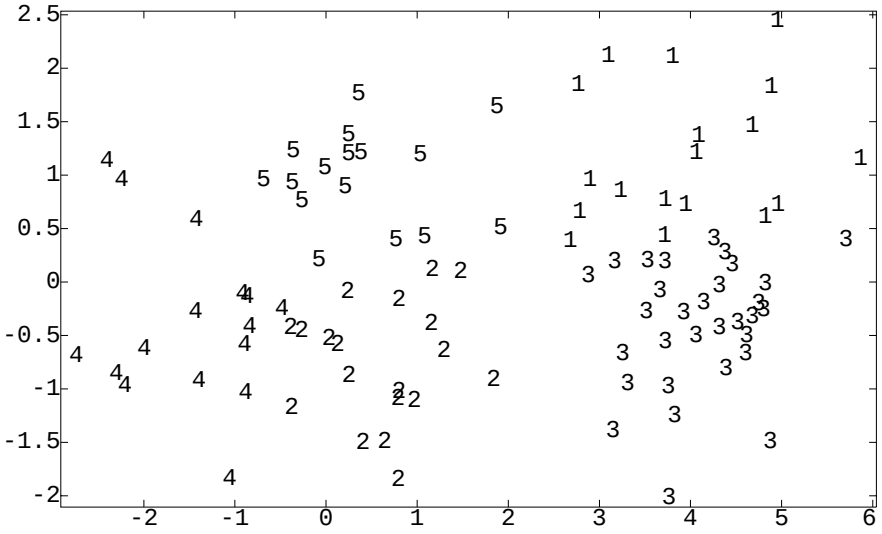
```
Cmd> out<-kmeans(X,kmax:10,standard:F)
Cluster analysis by reallocation of
 objects using Trace W criterion
Initial allocation is random
  k      Initial        Final Reallocations
 10        591.5       95.783      89
 10       95.783       78.358      22
 10       78.358       76.137       7
 10       76.137       65.895      12
 10       65.895       53.225      14
 10       53.225       51.035       7
 10       51.035       49.548       1
 10       49.548       49.408       2
 10       49.408       49.408       0


Cmd> chplot(X[,1],X[,2],out$classes,xaxis:F,yaxis:F)
```



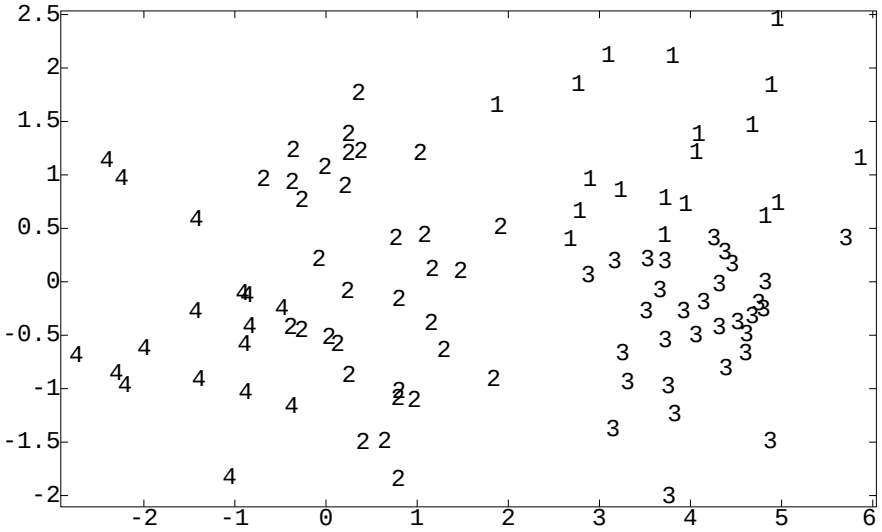
```
Cmd> out <- kmeans(X,kmax:10,kmin:2,standard:F)
Cluster analysis by reallocation of
 objects using Trace W criterion
Initial allocation is random
  k      Initial        Final Reallocations
 10       522.78       74.216      79
...
 10       47.936       47.936       0
Merging clusters 5 and 6; criterion = 51.891
...
  2       187.51       186.59       2
  2       186.59       186.59       0

Cmd> out$criterion
(1)  47.936    51.2  55.735   62.34  68.576
(6)  84.746  110.45  139.84  186.59

Cmd> tabs(,tclus,out$classes[,9])
(1,1)            2           48
(2,1)           49            1

Cmd> chplot(X[,1],X[,2],out$classes[,5],xaxis:F,yaxis:F)
```



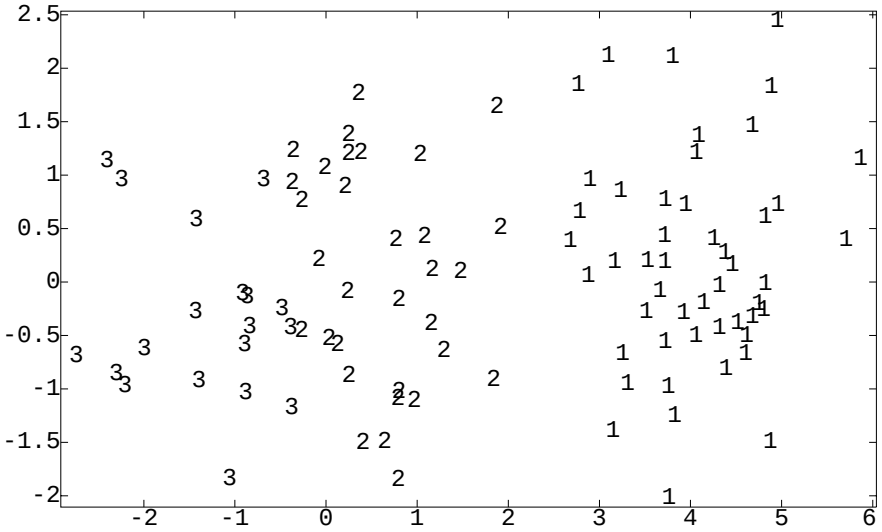
```
Cmd> chplot(X[,1],X[,2],out$classes[,6],xaxis:F,yaxis:F)
```



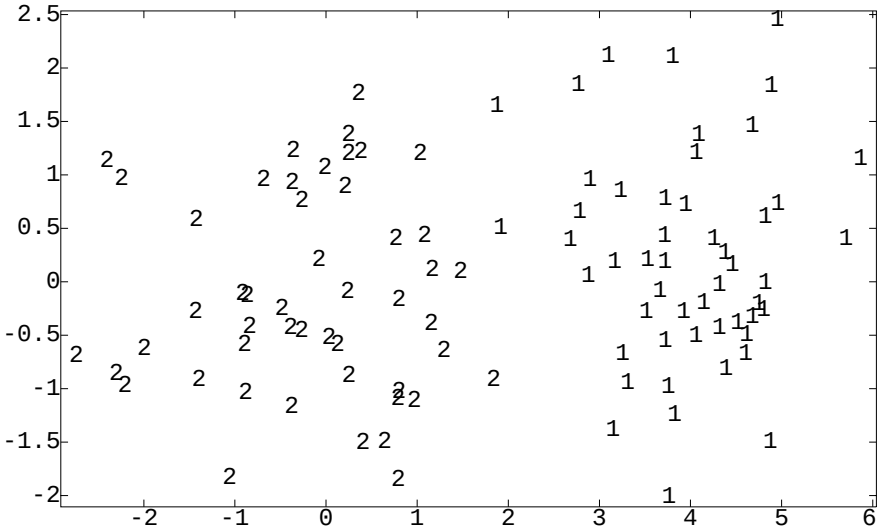
```
Cmd> chplot(X[,1],X[,2],out$classes[,7],xaxis:F,yaxis:F)
```



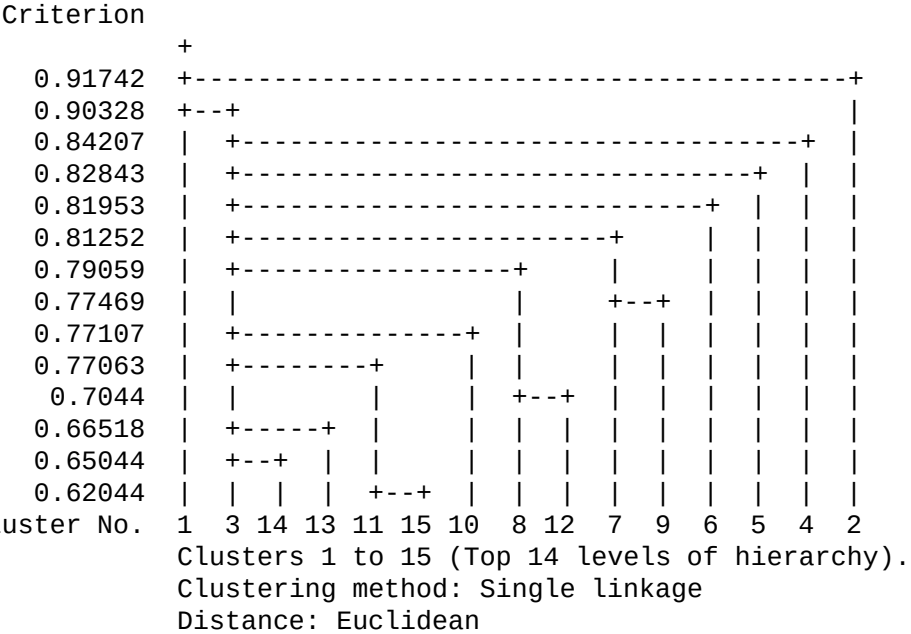
```
Cmd> chplot(X[,1],X[,2],out$classes[,8],xaxis:F,yaxis:F)
```



```
Cmd> chplot(X[,1],X[,2],out$classes[,9],xaxis:F,yaxis:F)
```

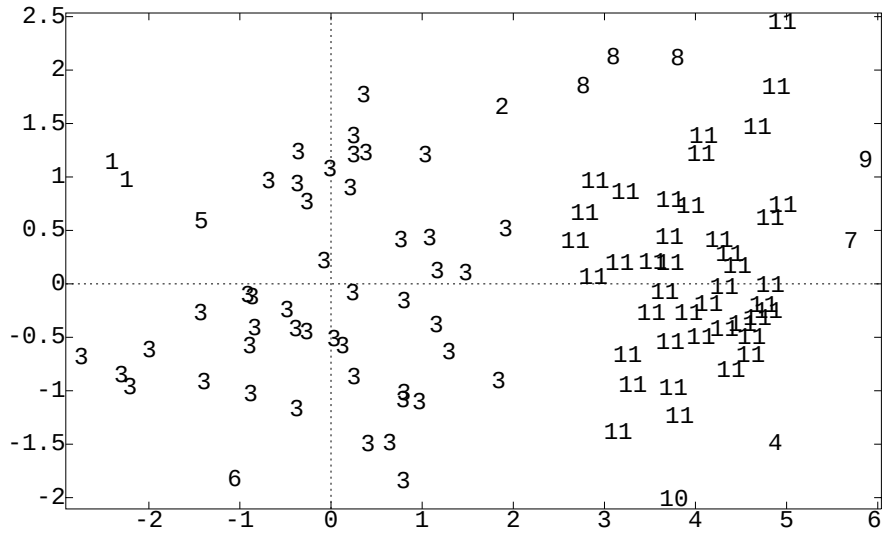


```
Cmd> cluster(X,standard:F,nclust:15,method:"single")
 Case  Number of Clusters
 No.   2  3  4  5  6  7  8  9 10 11 12 13 14 15
 ---- -- -- -- -- -- -- -- -- -- -- -- -- -- --
   1   1  1  1  1  1  1  1  1  1  1  1  1  1  1
   2   1  3  3  3  3  3  3  3  3  3  3  3  3  3
   3   1  3  3  3  3  3  3  3  3  3  3  3  3  3
...
 100   1  3  3  3  3  3  8  8  8  8 12 12 12 12
```

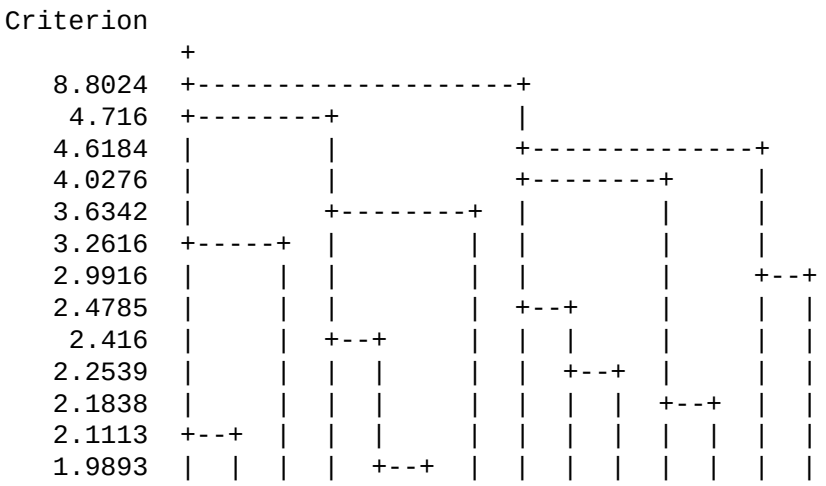


Clusters 1 to 15 (Top 14 levels of hierarchy).
Clustering method: Single linkage
Distance: Euclidean

```
Cmd> out <- cluster(X,standard:F,\
nclust:15,method:"single",keep:"classes")

Cmd> print(tabs(,tclus,out[,10]),format:"f2.0")
MATRIX:
(1,1)  2  1 45  0  1  1  0  0  0  0  0
(2,1)  0  0  1  1  0  0  1  3  1  1 42

Cmd> chplot(X[,1],X[,2],out[,10])
```

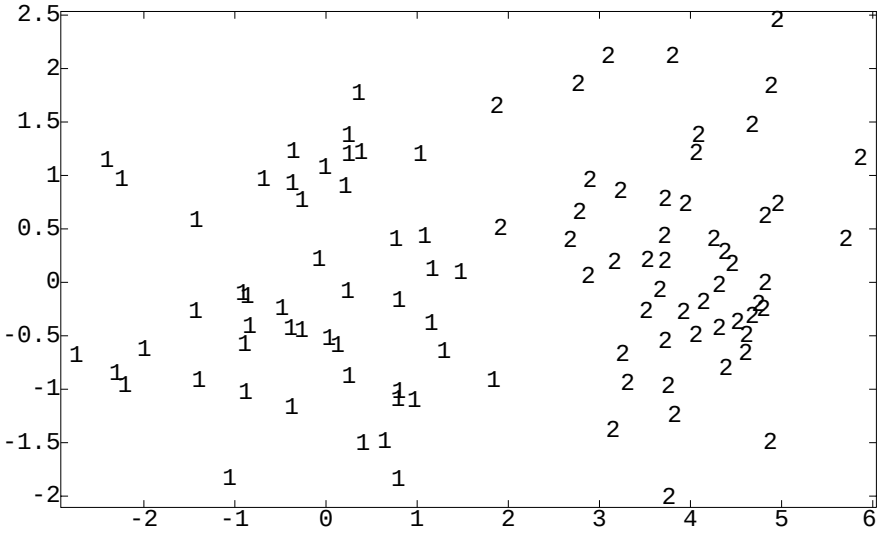


```
Cmd> out <- cluster(X,standard:F,nclust:15,\
method:"complete",keep:"classes")

Cmd>  cluster(X,standard:F,nclust:15,\
method:"complete")
```

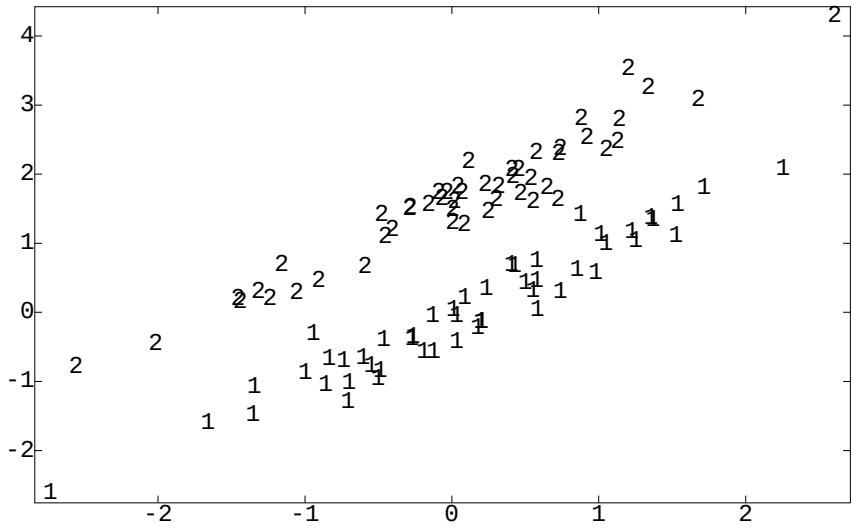


```
      1.8173  |  |  |  |  |  |  |  |  |  |  |  |  |  +--+
  Cluster No.  1 13  7  3 10 14  6  2  9 11  5 12  4  8 15
               Clusters 1 to 15 (Top 14 levels of hierarchy).
               Clustering method: Complete linkage
               Distance: Euclidean
```

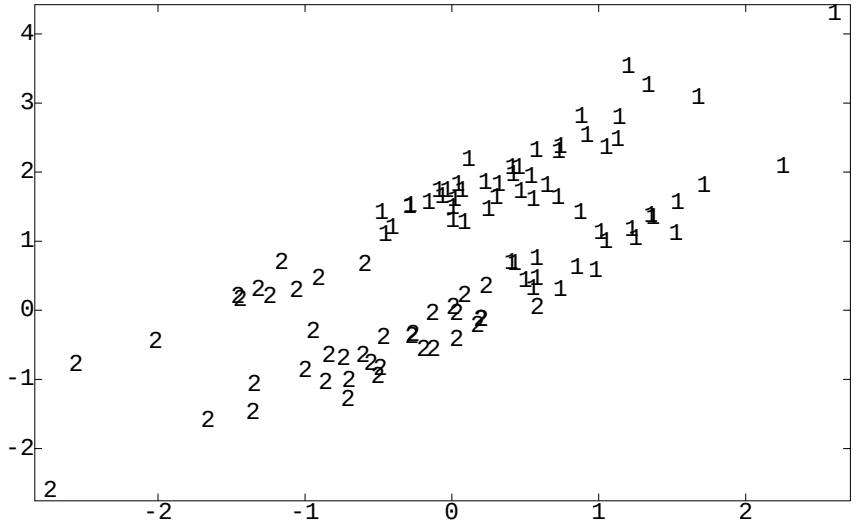
```
Cmd> chplot(X[,1],X[,2],out[,1],xaxis:F,yaxis:F)
```



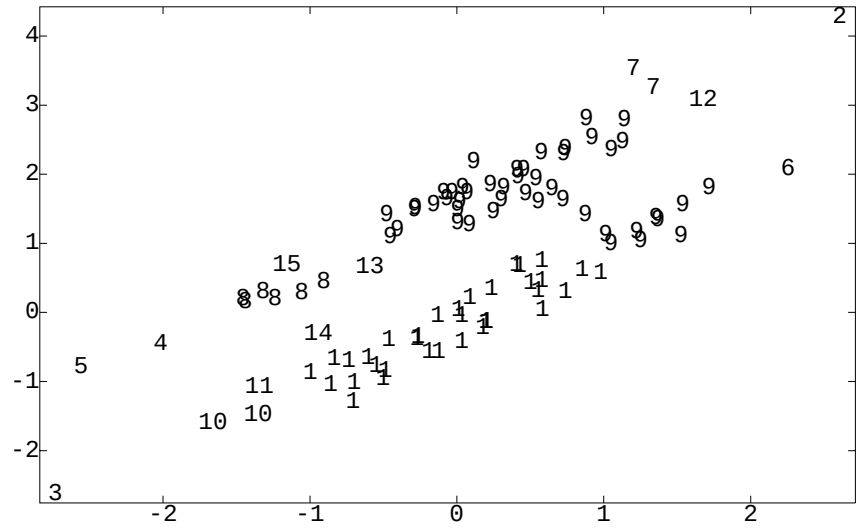

```
Cmd> X <- rnorm(100)*vector(1,1)'

Cmd> X <- X + matrix(rnorm(200),100)/5

Cmd> X[run(51,100),2] <- X[run(51,100),2]+1.5

Cmd> chplot(X[,1],X[,2],tclus,xaxis:F,yaxis:F)
```



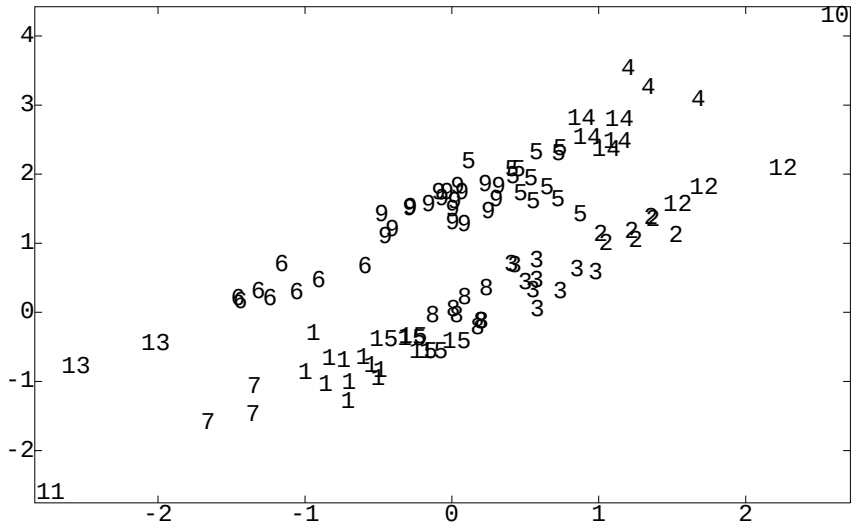
```
Cmd> out <- kmeans(X,standard:F,kmax:2)
Cluster analysis by reallocation of
 objects using Trace W criterion
Initial allocation is random
  k      Initial        Final Reallocations
  2       248.11       107.44      51
  2       107.44       101.83       8
  2       101.83       101.11       2
  2       101.11       100.86       1
  2       100.86       100.86       0

Cmd> chplot(X[,1],X[,2],out$classes,xaxis:F,yaxis:F)
```



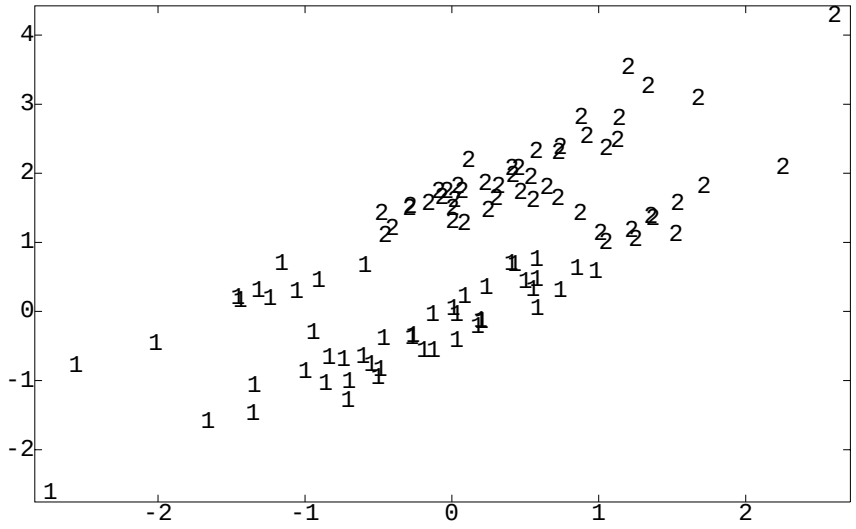
```
Cmd> out <- cluster(X,standard:F,\
method:"single",nclust:15,keep:"classes")

Cmd> tabs(,out[,14])
 (1)  34   1   1   1   1
 (6)   1   2   6  46   2
(11)   1   1   1   1   1

Cmd> chplot(X[,1],X[,2],out[,14],xaxis:F,yaxis:F)
```



```
Cmd> out <- cluster(X,standard:F,\
method:"ward",nclust:15,keep:"classes")

Cmd> chplot(X[,1],X[,2],out[,14],xaxis:F,yaxis:F)
```

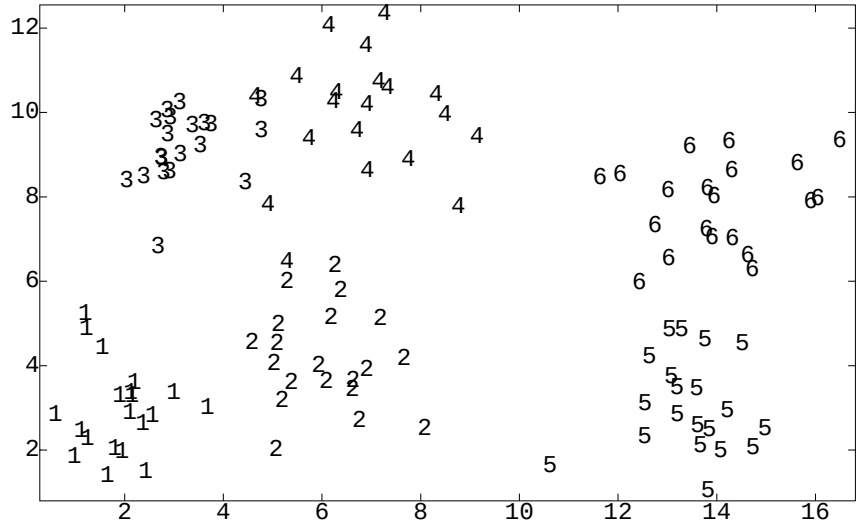


```
Cmd> chplot(X[,1],X[,2],out[,1],xaxis:F,yaxis:F)
```

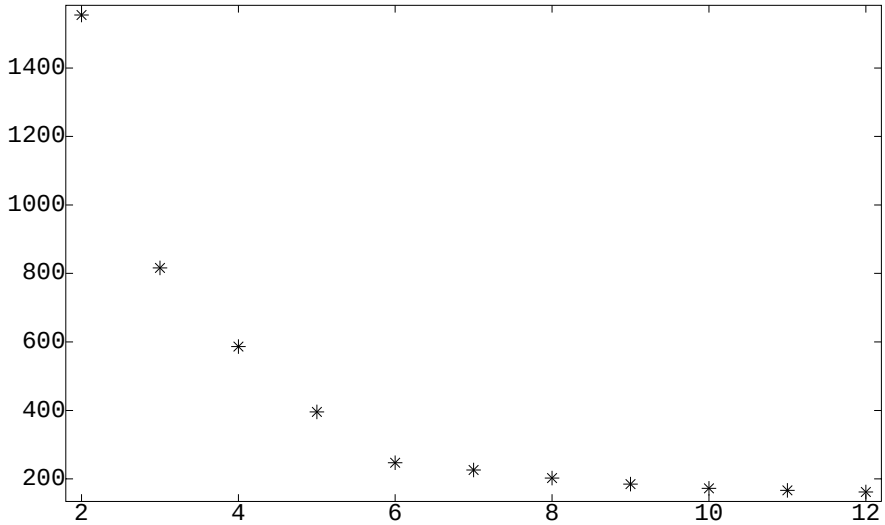


How many clusters? That's a very good question.
Usually we try several different numbers of clusters and plot the criterion. The criterion shouldn't change much as long as we are joining "subclusters". It will probably jump up when we join two truly different clusters.
This guideline works for hierarchical as well as k-means.

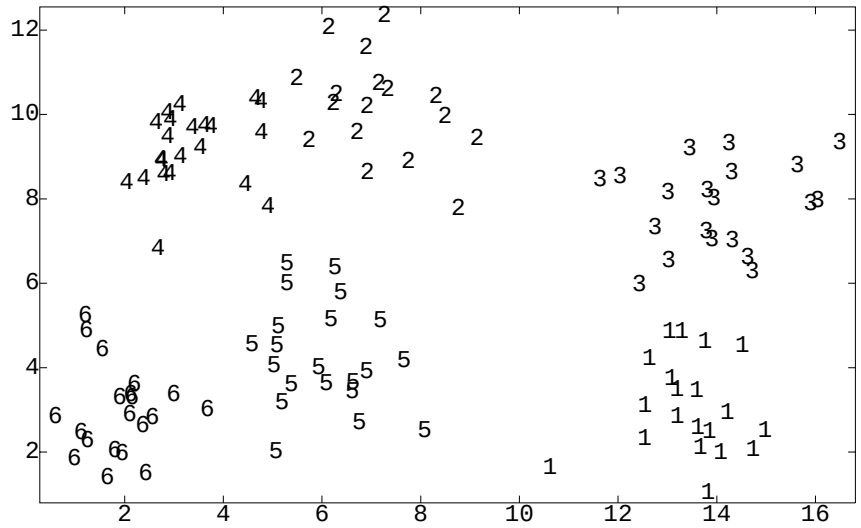
```
Cmd> X <- matrix(rnorm(240),120)

Cmd> X[,1] <- X[,1] + vector(2,6,3,7,13,14)[tclus]

Cmd> X[,2] <- X[,2] + vector(3,4,9,10,3,8)[tclus]

Cmd> chplot(X[,1],X[,2],tclus)
```



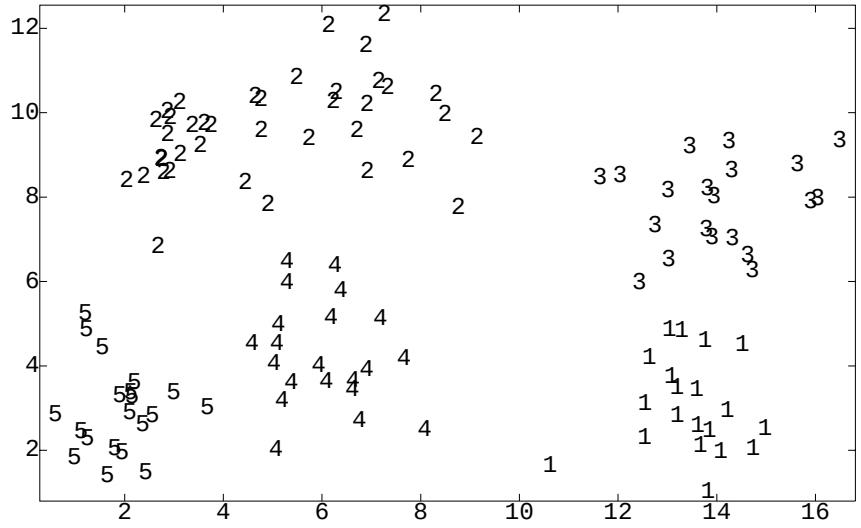
```
Cmd> out <- kmeans(X,standard:F,kmax:12,kmin:2)
Cluster analysis by reallocation of objects using Trace W criterion
Initial allocation is random
...

Cmd> out$criterion
 (1) 161.84 166.36 172.14 184.88 202.62
 (6) 226.16 247.46 395.49 586.43  816.1
(11)   1555

Cmd> plot(run(2,12),reverse(out$criterion))
```



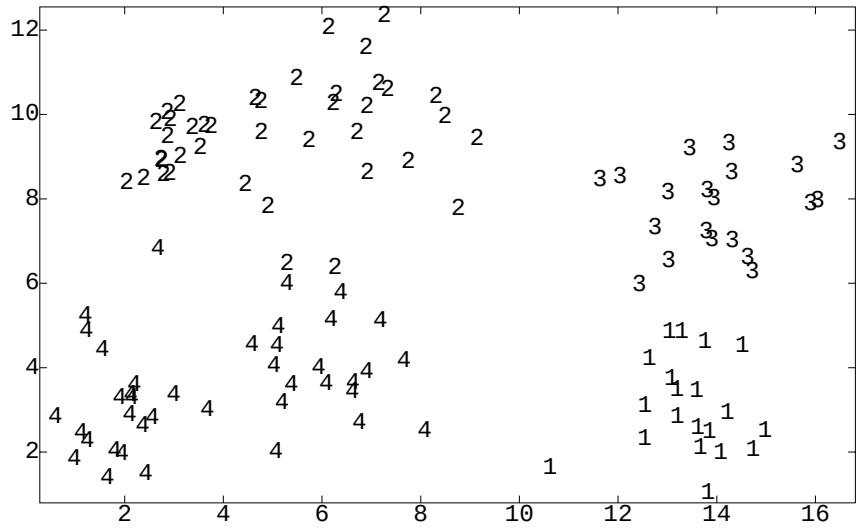
```
Cmd> chplot(X[,1],X[,2],out$classes[,7])
```



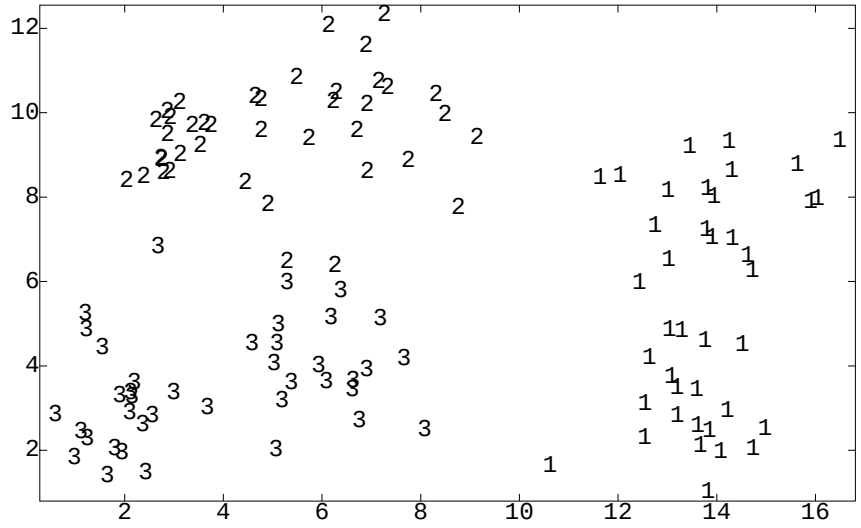
```
Cmd> chplot(X[,1],X[,2],out$classes[,8])
```



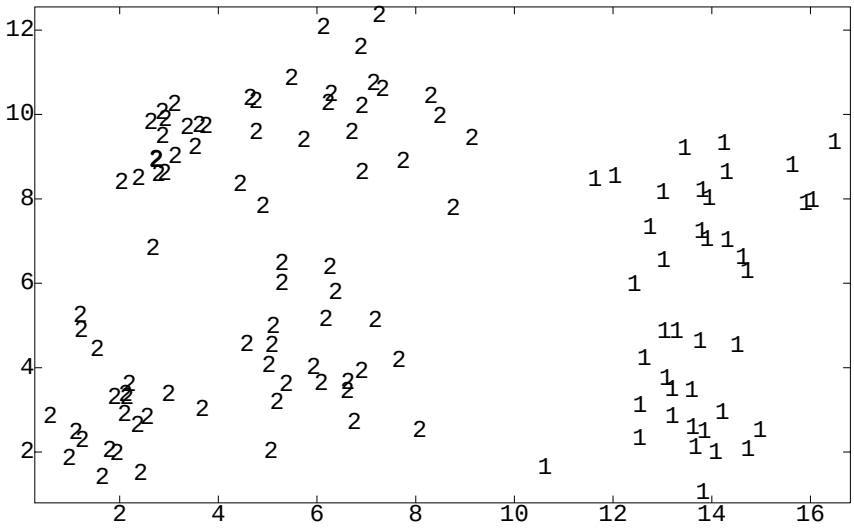
```
Cmd> chplot(X[,1],X[,2],out$classes[,9])
```



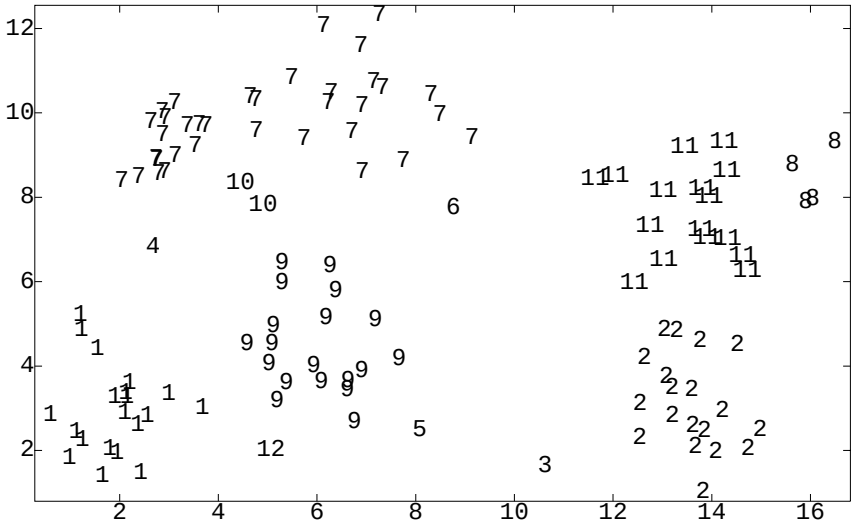
```
Cmd> chplot(X[,1],X[,2],out$classes[,10])
```



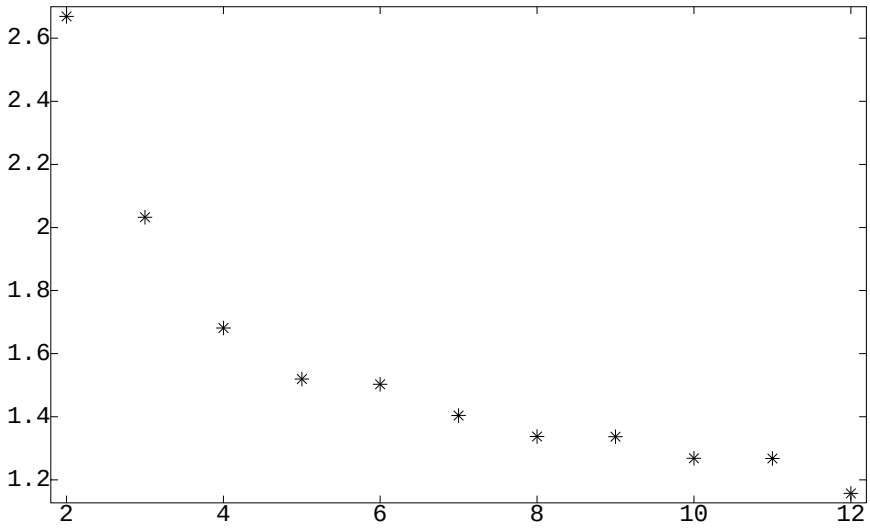
```
Cmd> chplot(X[,1],X[,2],out$classes[,11])
```



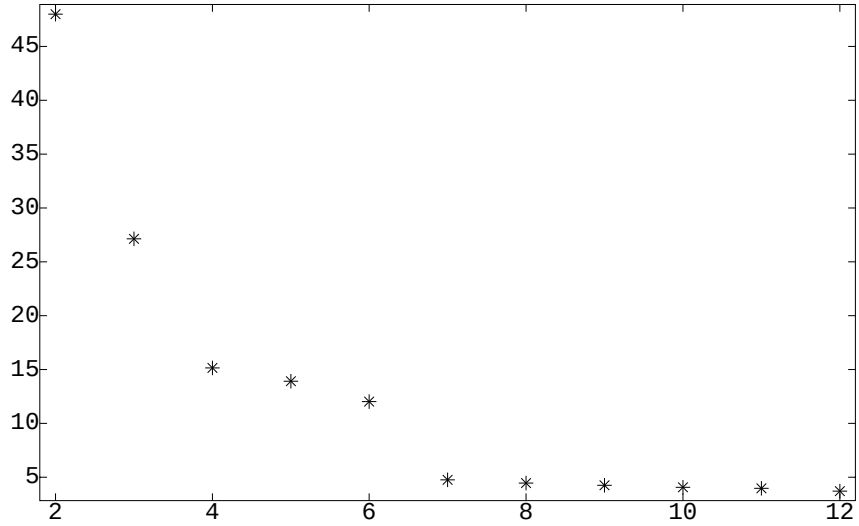
```
Cmd> out <- cluster(X,standard:F,nclust:12,\
method:"single",keep:"all")

Cmd> tabs(,out$classes[,11])
 (1)    20    19     1     1     1
 (6)     1    35     4    19     2
(11)    16     1

Cmd> chplot(X[,1],X[,2],out$classes[,11])
```



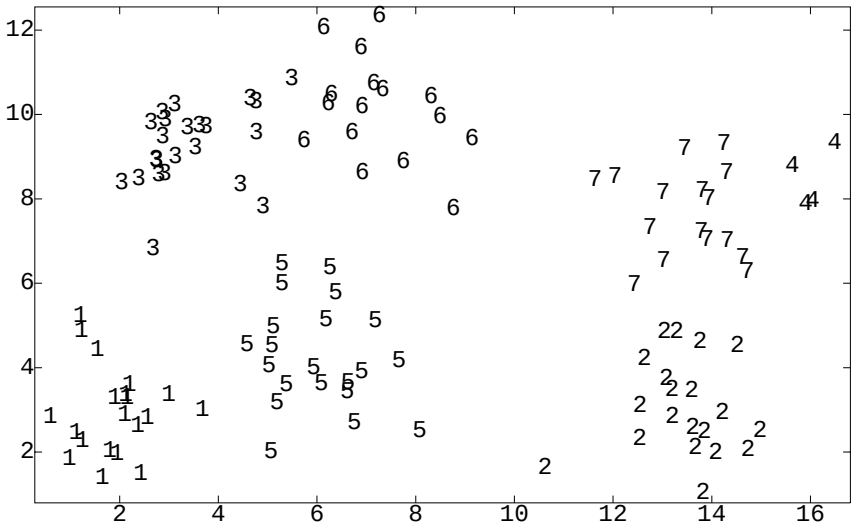
```
Cmd> plot(run(2,12),out$criterion)
```




```
Cmd> out <- cluster(X,standard:F,nclust:12,\
method:"ward",keep:"all")

Cmd> plot(run(2,12),out$criterion)
```




```
Cmd> chplot(X[,1],X[,2],out$classes[,6])
```


12
10
8
6
4
2
2
4
6
8
10
12
14
16