# 08 Unsupervised Learning Homework

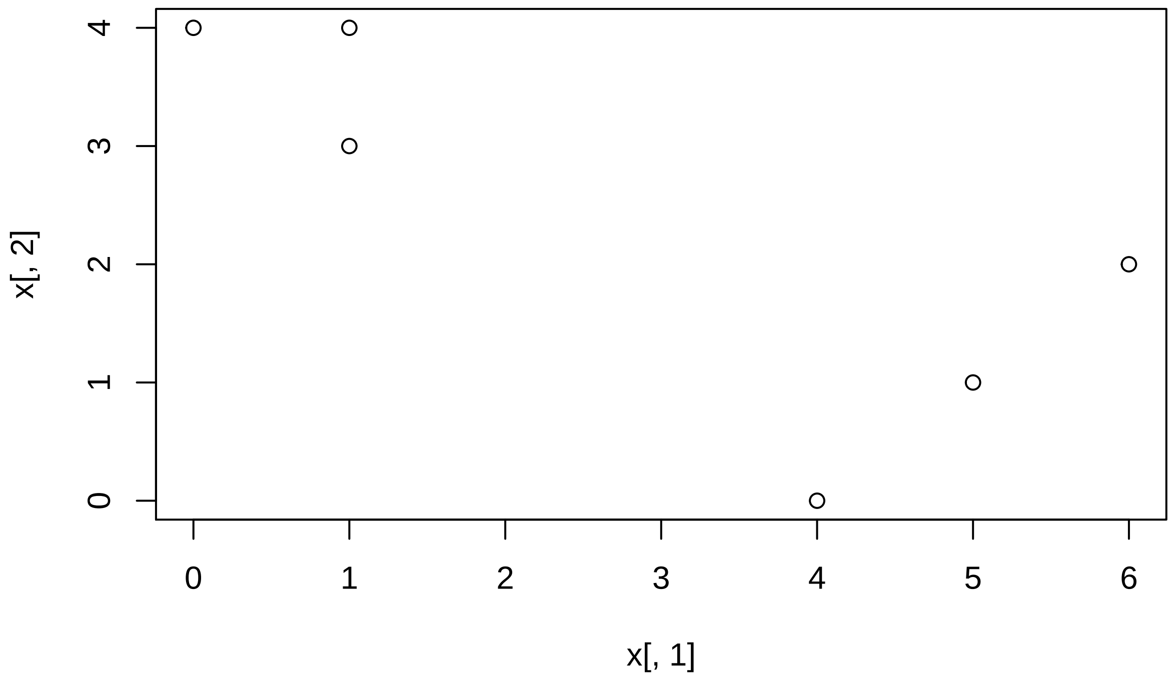
*Xinci Chen*

*3/16/2020*

### 10.3

#### (a)

```
x <- cbind(c(1, 1, 0, 5, 6, 4), c(4, 3, 4, 1, 2, 0))
plot(x[,1], x[,2])
```

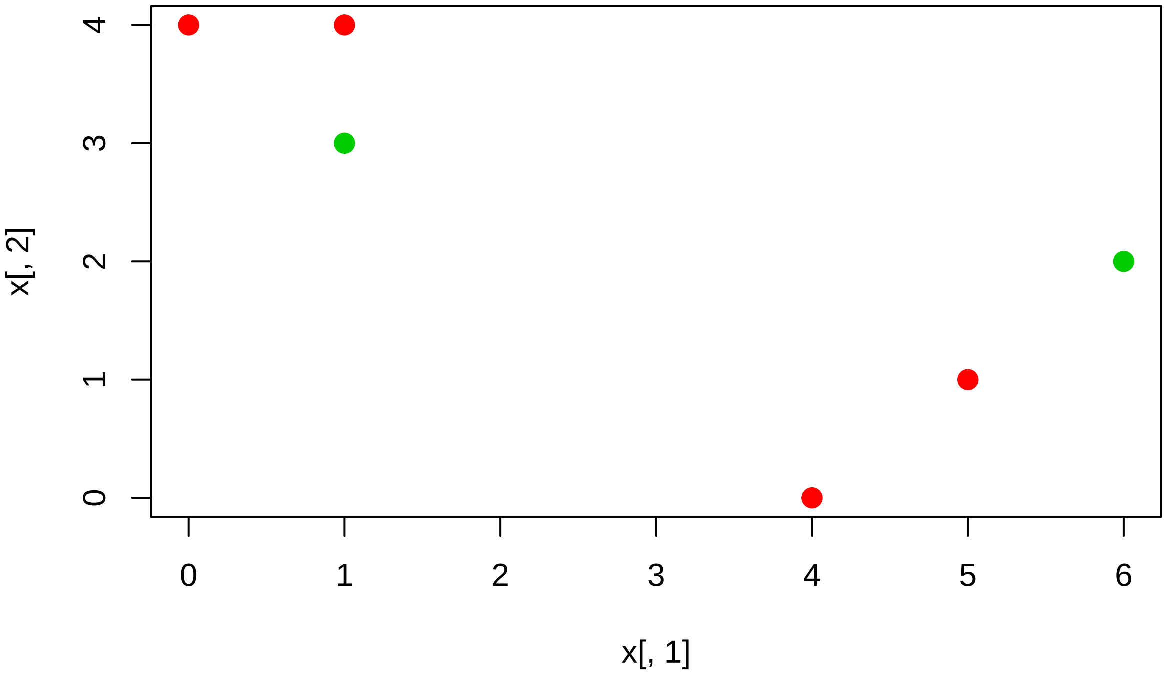


#### (b)

```
set.seed(1)
labels <- sample(2, nrow(x), replace = T)
labels
```

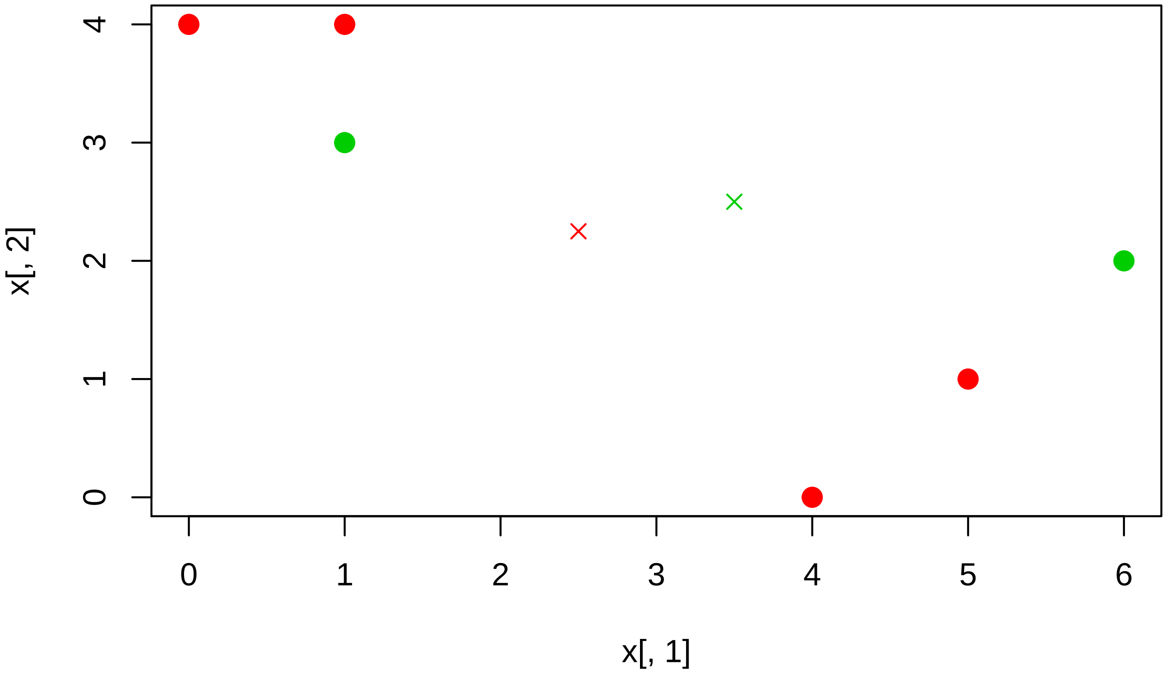
```
## [1] 1 2 1 1 2 1
```

```
plot(x[, 1], x[, 2], col = (labels + 1), pch = 20, cex = 2)
```



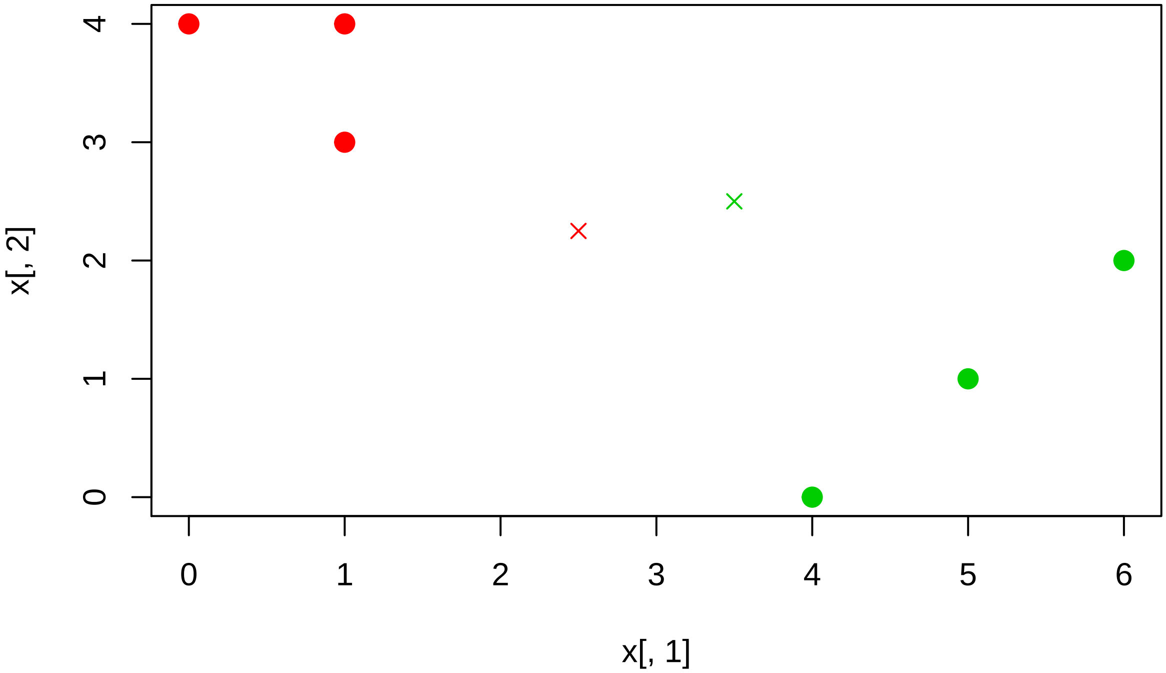
**(c)**

```
centroid1 <- c(mean(x[labels == 1, 1]), mean(x[labels == 1, 2]))
centroid2 <- c(mean(x[labels == 2, 1]), mean(x[labels == 2, 2]))
plot(x[,1], x[,2], col=(labels + 1), pch = 20, cex = 2)
points(centroid1[1], centroid1[2], col = 2, pch = 4)
points(centroid2[1], centroid2[2], col = 3, pch = 4)
```



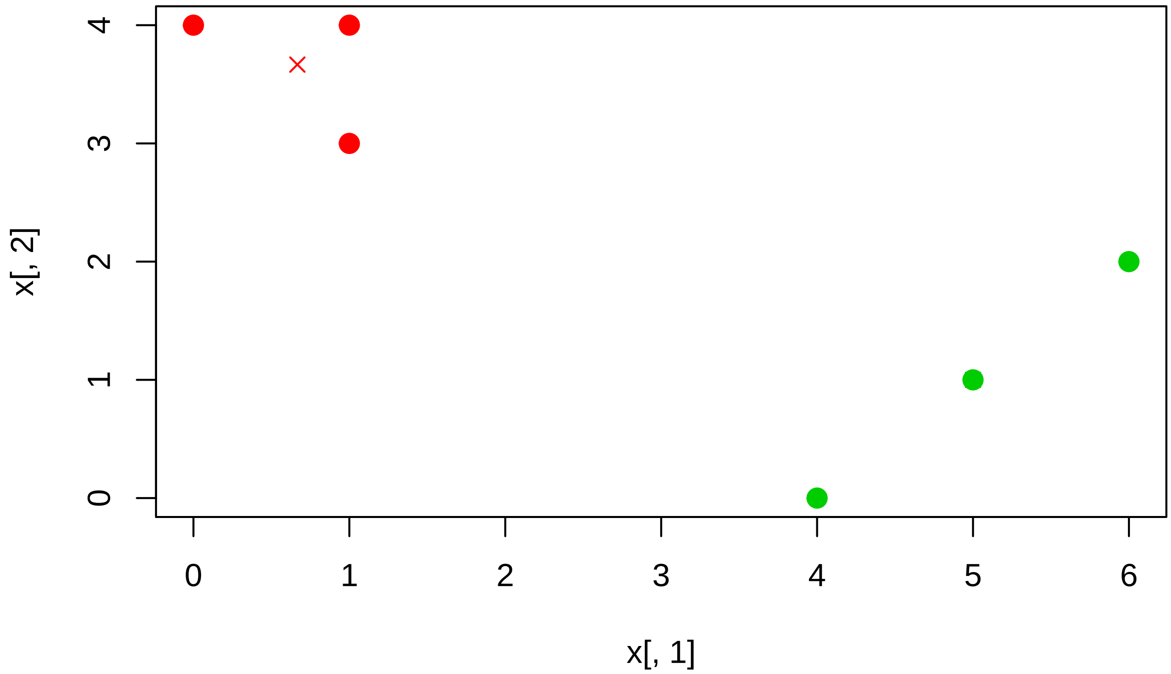
### (d)

```
labels <- c(1, 1, 1, 2, 2, 2)
plot(x[, 1], x[, 2], col = (labels + 1), pch = 20, cex = 2)
points(centroid1[1], centroid1[2], col = 2, pch = 4)
points(centroid2[1], centroid2[2], col = 3, pch = 4)
```



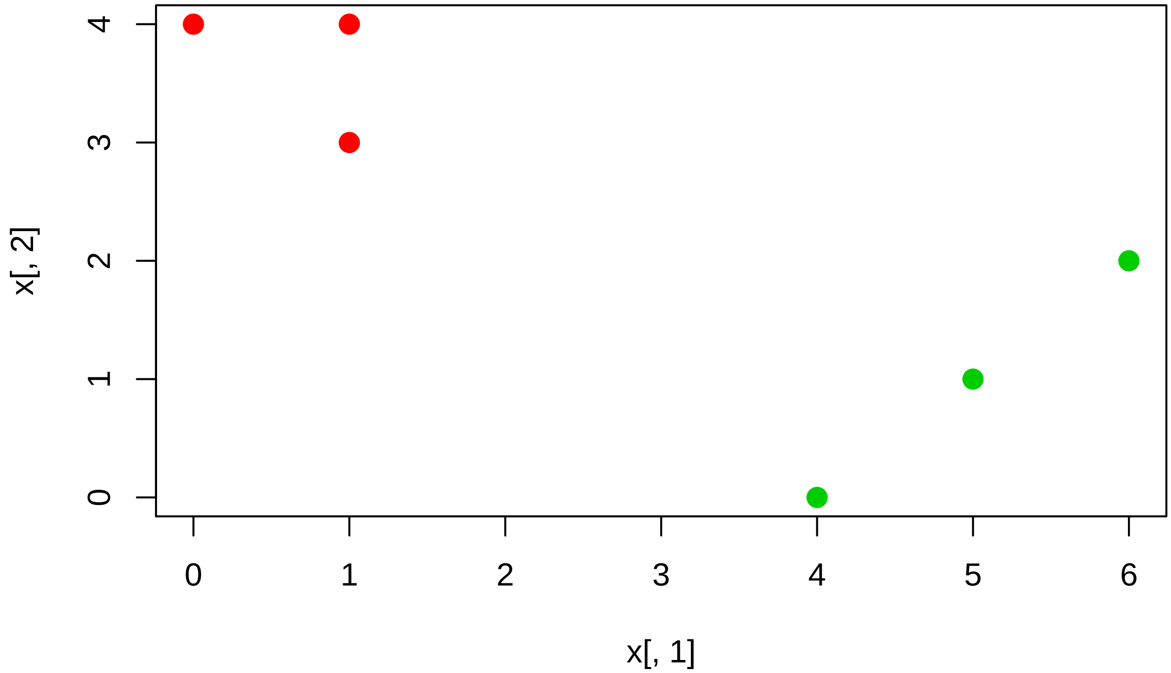
### (e)

```
centroid1 <- c(mean(x[labels == 1, 1]), mean(x[labels == 1, 2]))
centroid2 <- c(mean(x[labels == 2, 1]), mean(x[labels == 2, 2]))
plot(x[,1], x[,2], col=(labels + 1), pch = 20, cex = 2)
points(centroid1[1], centroid1[2], col = 2, pch = 4)
points(centroid2[1], centroid2[2], col = 3, pch = 4)
```



##(f)

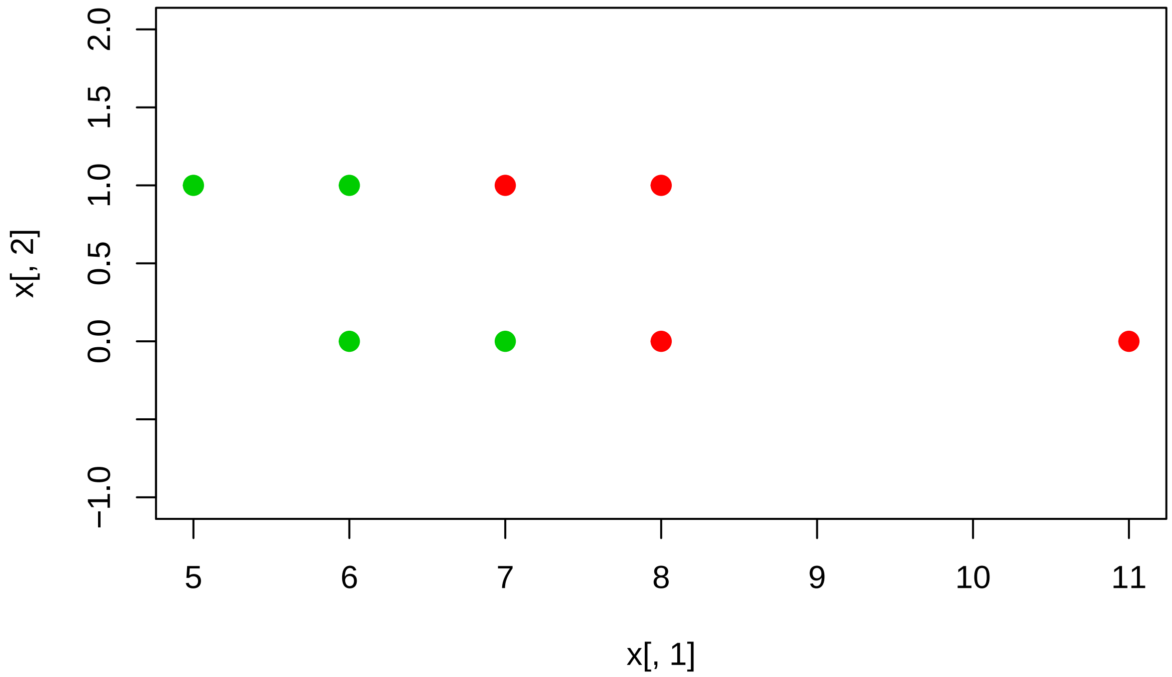
```
plot(x[, 1], x[, 2], col=(labels + 1), pch = 20, cex = 2)
```



### 10.5

```
socks <- c(8, 11, 7, 6, 5, 6, 7, 8)
computers <- c(0, 0, 0, 0, 1, 1, 1, 1)
x <- cbind(socks, computers)
labels <- c(1, 1, 2, 2, 2, 2, 1, 1)
```

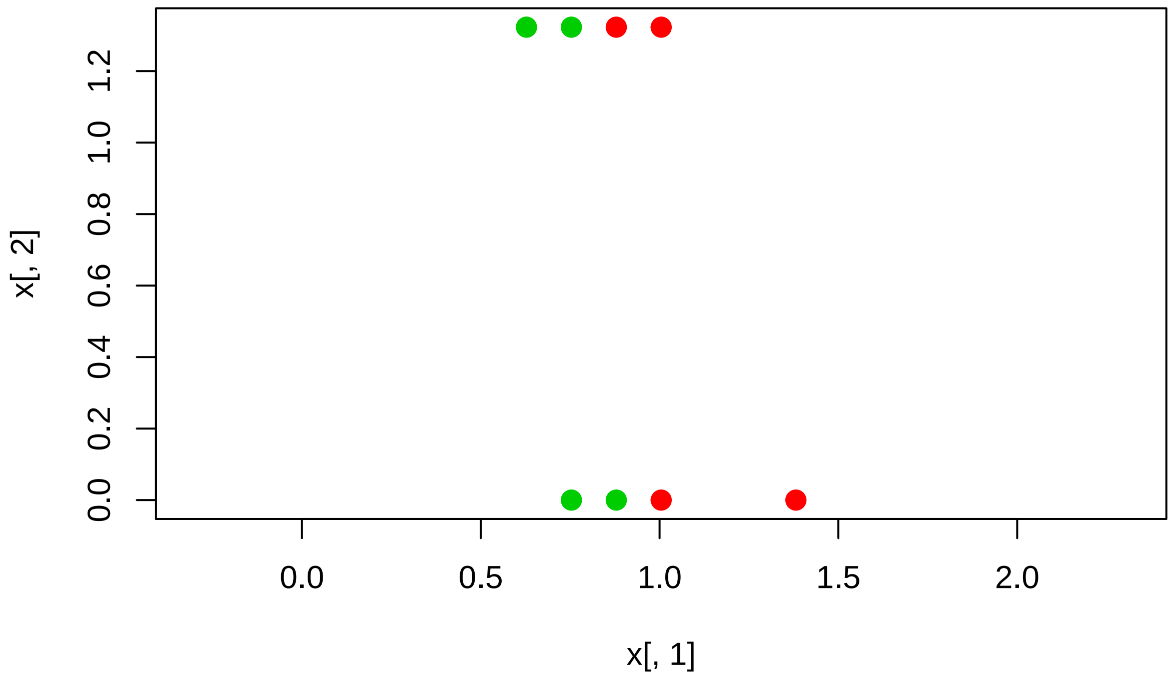
```
plot(x[, 1], x[, 2], col=(labels + 1), pch = 20, cex = 2, asp = 1)
```



```
x <- cbind(scale(socks, center = FALSE), scale(computers, center = FALSE))
sd(computers)
```

```
## [1] 0.5345225
```

```
labels <- c(1, 1, 2, 2, 2, 2, 1, 1)
plot(x[, 1], x[, 2], col=(labels + 1), pch = 20, cex = 2, asp = 1)
```



### 10.6

### (a)

90% of the variance in the data is not contained in the first principal component.

### (c)

```r
set.seed(1)
Control <- matrix(rnorm(50 * 1000), ncol = 50)
Treatment <- matrix(rnorm(50 * 1000), ncol = 50)
X <- cbind(Control, Treatment)
X[1, ] <- seq(-18, 18 - .36, .36) # linear trend in one dimension
pr.out <- prcomp(scale(X))
summary(pr.out)$importance[, 1]
```

```
##     Standard deviation Proportion of Variance  Cumulative Proportion 
##               3.148148               0.099110               0.099110
```

```r
X <- rbind(X, c(rep(10, 50), rep(0, 50)))
pr.out <- prcomp(scale(X))
summary(pr.out)$importance[, 1]
```

```
##     Standard deviation Proportion of Variance  Cumulative Proportion 
##               3.397839               0.115450               0.115450
```

### 10.8

### (a)

```r
pr.out <- prcomp(USArrests, scale = TRUE)
pr.var <- pr.out$sdev^2
pve <- pr.var / sum(pr.var)
sum(pr.var)
```

```
## [1] 4
```

### (b)

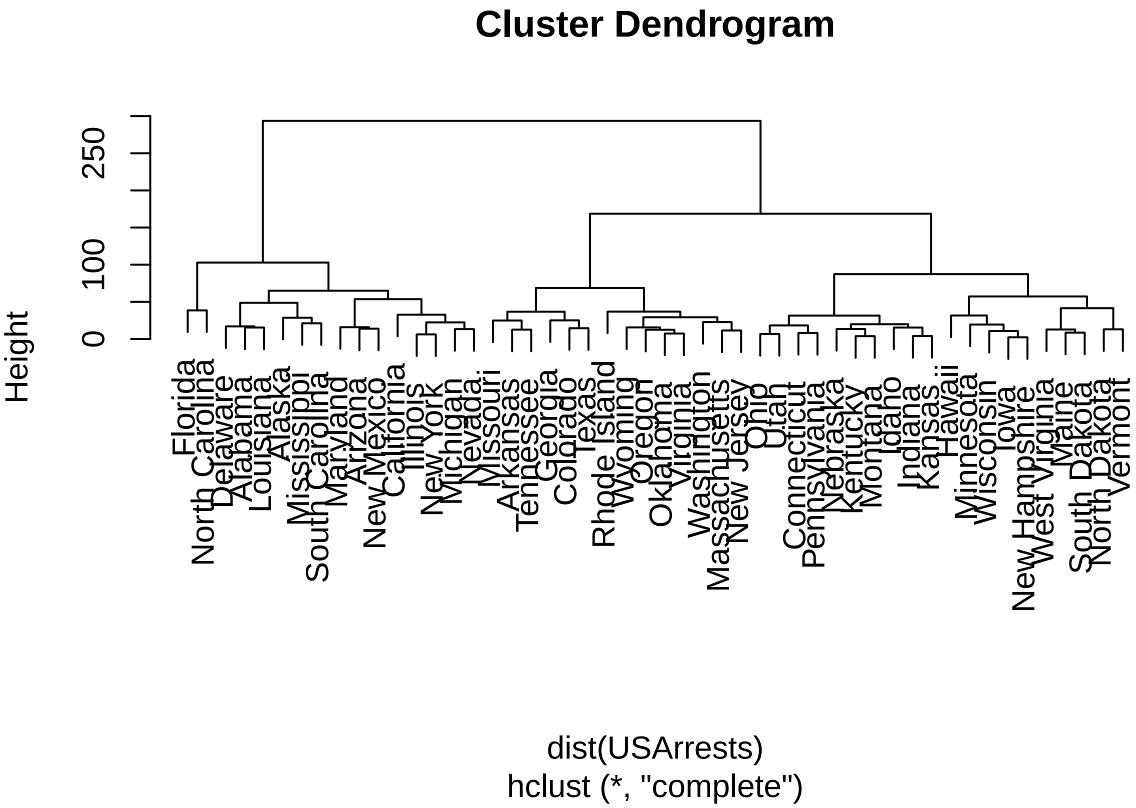
```r
loadings <- pr.out$rotation
USArrests2 <- scale(USArrests)
sumvar <- sum(apply(as.matrix(USArrests2)^2, 2, sum))
apply((as.matrix(USArrests2) %*% loadings)^2, 2, sum) / sumvar
```

```
##        PC1        PC2        PC3        PC4 
## 0.62006039 0.24744129 0.08914080 0.04335752
```

### 10.9

### (a)

```r
set.seed(2)
hc.complete <- hclust(dist(USArrests), method = "complete")
plot(hc.complete)
```



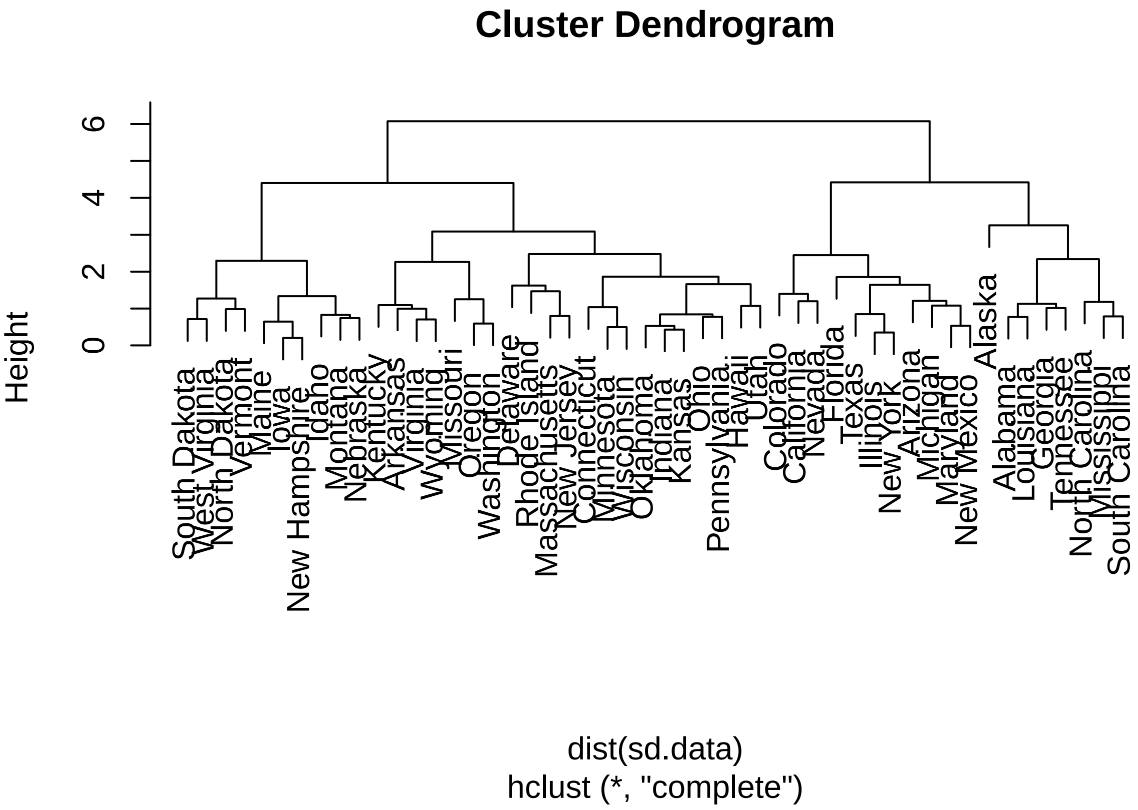
**(b)**

```
cutree(hc.complete, 3)
```

```
##        Alabama         Alaska        Arizona       Arkansas     California 
##              1              1              1              2              1 
##       Colorado    Connecticut       Delaware        Florida        Georgia 
##              2              3              1              1              2 
##         Hawaii          Idaho       Illinois        Indiana           Iowa 
##              3              3              1              3              3 
##         Kansas       Kentucky      Louisiana          Maine       Maryland 
##              3              3              1              3              1 
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri 
##              2              1              3              1              2 
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey 
##              3              3              1              3              2 
##     New Mexico       New York North Carolina   North Dakota           Ohio 
##              1              1              1              3              3 
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina 
##              2              2              3              2              1 
##   South Dakota      Tennessee          Texas           Utah        Vermont 
##              3              2              2              3              3 
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming 
##              2              2              3              3              2
```

### (c)

```
sd.data <- scale(USArrests)
hc.complete.sd <- hclust(dist(sd.data), method = "complete")
plot(hc.complete.sd)
```



### (d)

```
cutree(hc.complete.sd, 3)
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              1              2              3              2
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              3              3              2              1
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              3              3              2              3              3
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              3              3              1              3              2
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              3              2              3              1              3
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              3              3              2              3              3
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              2              2              1              3              3
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              3              3              3              3              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              3              1              2              3              3
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming 
##              3              3              3              3              3
```

```
table(cutree(hc.complete, 3), cutree(hc.complete.sd, 3))
```

```
## 
##       1  2  3
##   1   6  9  1
##   2   2  2 10
##   3   0  0 20
```

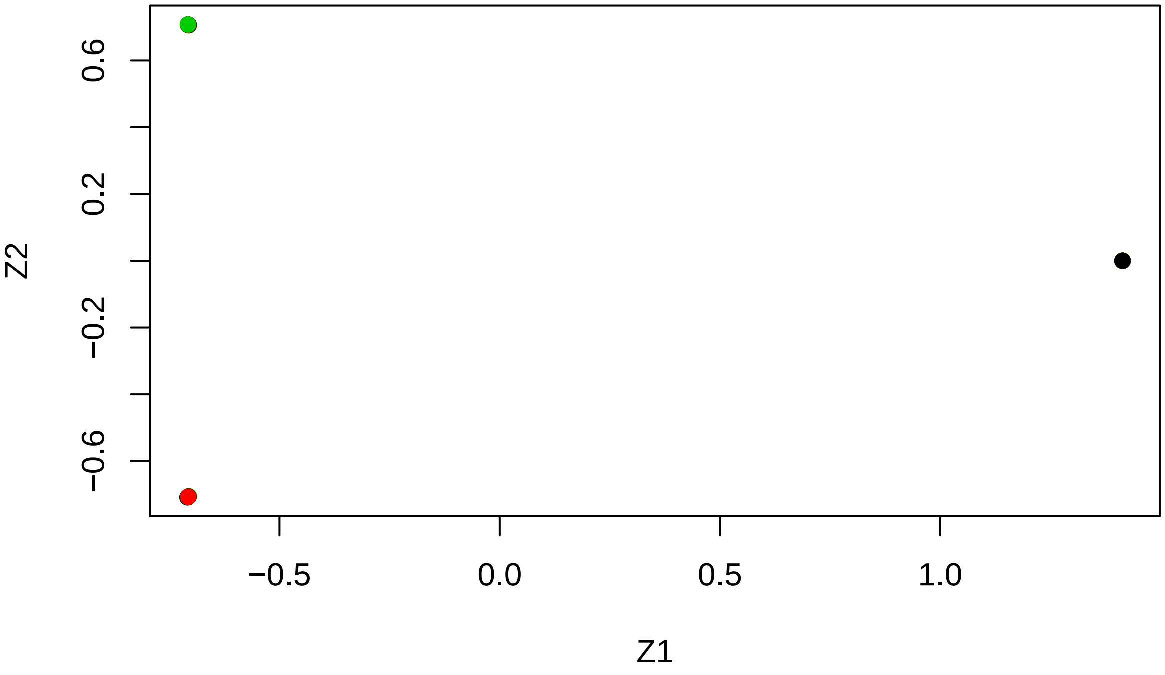
### 10.10

#### (a)

```
set.seed(2)
x <- matrix(rnorm(20 * 3 * 50, mean = 0, sd = 0.001), ncol = 50)
x[1:20, 2] <- 1
x[21:40, 1] <- 2
x[21:40, 2] <- 2
x[41:60, 1] <- 1
true.labels <- c(rep(1, 20), rep(2, 20), rep(3, 20))
```

#### (b)

```
pr.out <- prcomp(x)
plot(pr.out$x[, 1:2], col = 1:3, xlab = "Z1", ylab = "Z2", pch = 19)
```



#### (c)

```
km.out <- kmeans(x, 3, nstart = 20)
table(true.labels, km.out$cluster)
```

```
## 
## true.labels  1  2  3
##           1  0  0 20
##           2 20  0  0
##           3  0 20  0
```

## (d)

```r
km.out <- kmeans(x, 2, nstart = 20)
table(true.labels, km.out$cluster)
```

```
## 
## true.labels  1  2
##           1 20  0
##           2  0 20
##           3 20  0
```

## (e)

```r
km.out <- kmeans(x, 4, nstart = 20)
table(true.labels, km.out$cluster)
```

```
## 
## true.labels  1  2  3  4
##           1 11  9  0  0
##           2  0  0 20  0
##           3  0  0  0 20
```

## (f)

```r
km.out <- kmeans(pr.out$x[, 1:2], 3, nstart = 20)
table(true.labels, km.out$cluster)
```

```
## 
## true.labels  1  2  3
##           1  0  0 20
##           2  0 20  0
##           3 20  0  0
```

## (g)

```r
km.out <- kmeans(scale(x), 3, nstart = 20)
table(true.labels, km.out$cluster)
```

```
## 
## true.labels  1  2  3
##           1  9  2  9
##           2  2 18  0
##           3  7  1 12
```