Multivariate Analysis for the Behavioral Sciences, Second Edition (Chapman and Hall/CRC, 2019) **Solutions to Exercises of Chapter 17: Cluster Analysis**

Kimmo Vehkalahti and Brian S. Everitt

19 December 2018

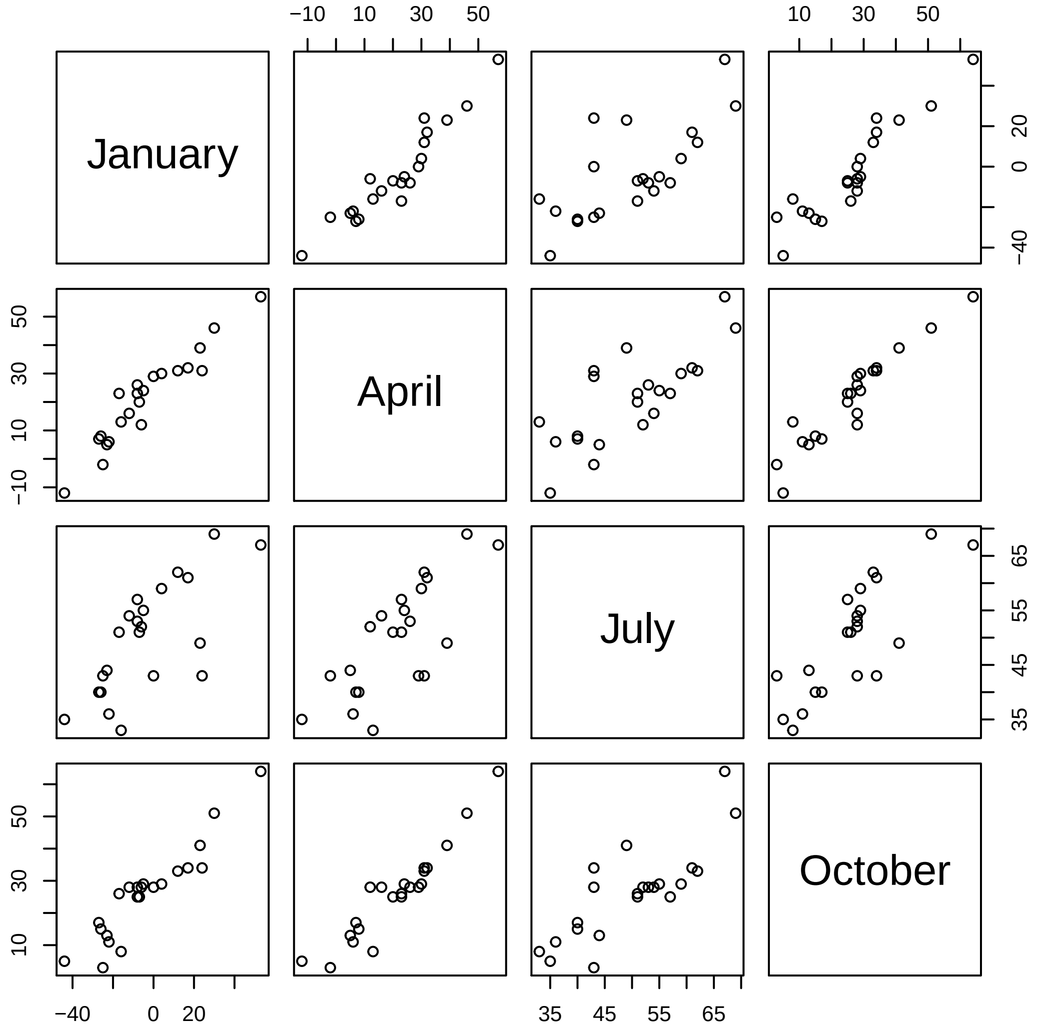
Solutions

Exercise 17.3

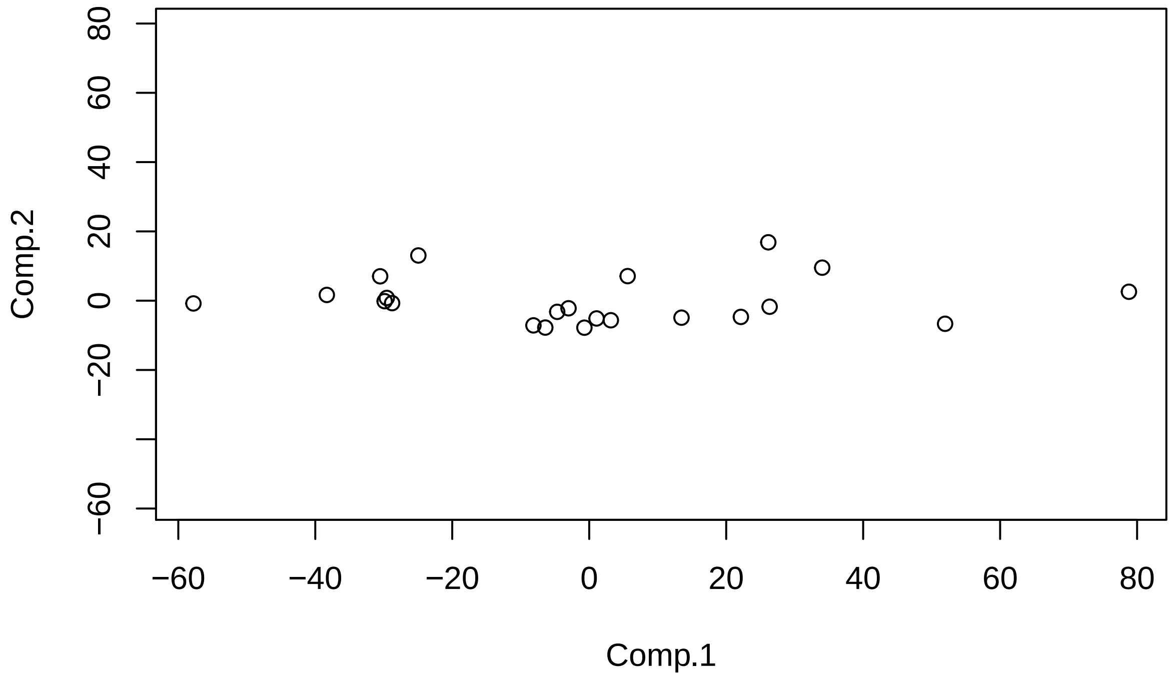
```
lowtemp <- structure(
  c(-8, -7, -44, -12, -27, 4, -25, -8, 53, 12, -22, 23, 30, -17, -6, -23, 17, -26, -16, 24,
    0, -5, 26, 20, -12, 16, 7, 30, -2, 23, 57, 31, 6, 39, 46, 23, 12, 5, 32, 8, 13, 31, 29,
    24, 53, 51, 35, 54, 40, 59, 43, 57, 67, 62, 36, 49, 69, 51, 52, 44, 61, 40, 33, 43, 43,
    55, 28, 25, 5, 28, 17, 29, 3, 25, 64, 33, 11, 41, 51, 26, 28, 13, 34, 15, 8, 34, 28, 29),
.Dim = c(22L, 4L), .Dimnames = list(c("Atlanta", "Baltimore", "Bismark", "Boston", "Chicago",
                   "Dallas", "Denver", "El Paso", "Honolulu", "Houston", "Juneau", "Los Angeles",
                   "Miami", "Nashville", "New York", "Omaha", "Phoenix", "Portland", "Reno",
                   "San Francisco", "Seattle", "Washington"),
                 c("January", "April", "July", "October")))
lowtemp
```

```
##               January April July October
## Atlanta            -8    26   53      28
## Baltimore          -7    20   51      25
## Bismark           -44   -12   35       5
## Boston            -12    16   54      28
## Chicago           -27     7   40      17
## Dallas              4    30   59      29
## Denver            -25    -2   43       3
## El Paso            -8    23   57      25
## Honolulu           53    57   67      64
## Houston            12    31   62      33
## Juneau            -22     6   36      11
## Los Angeles        23    39   49      41
## Miami              30    46   69      51
## Nashville         -17    23   51      26
## New York           -6    12   52      28
## Omaha             -23     5   44      13
## Phoenix            17    32   61      34
## Portland          -26     8   40      15
## Reno              -16    13   33       8
## San Francisco      24    31   43      34
## Seattle             0    29   43      28
## Washington         -5    24   55      29
```

```
pairs(lowtemp)
```



```
lowtemp_pc <- princomp(lowtemp)
xlim <- range(lowtemp_pc$scores[, 1])
plot(lowtemp_pc$scores[, 1:2], ylim = xlim)
```



```
# possibly 2 or three clusters?
```

```r
lowtemp_km2 <- kmeans(lowtemp, 2)
lowtemp_km2
```

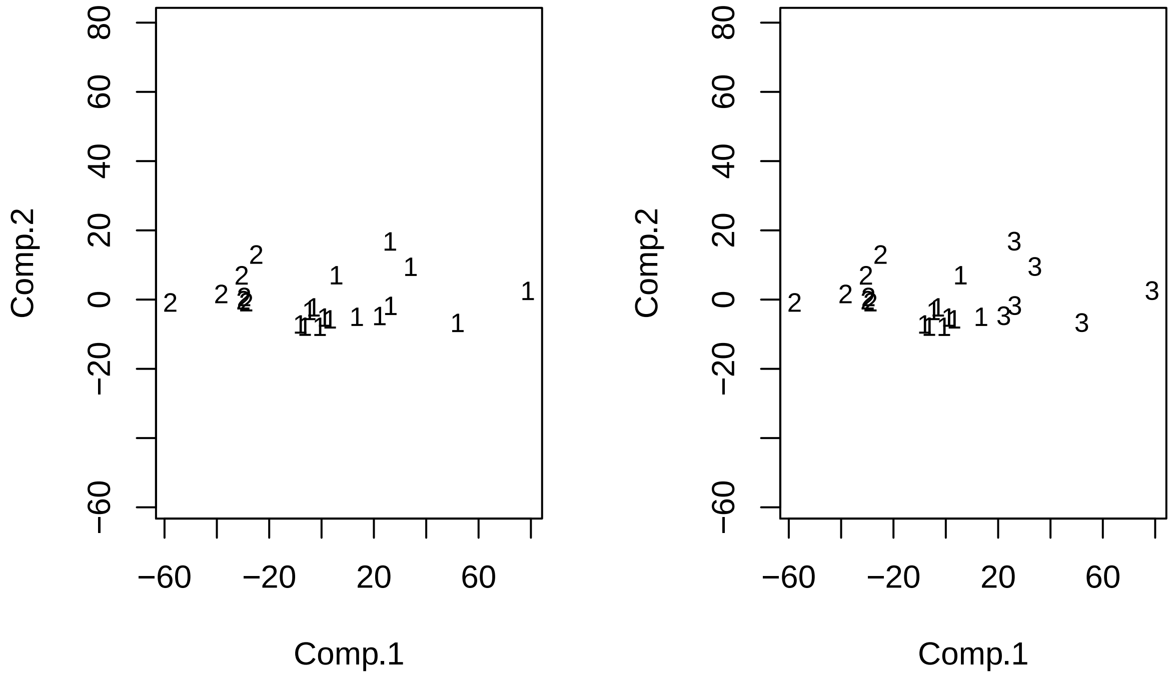
```
## K-means clustering with 2 clusters of sizes 15, 7
##
## Cluster means:
##      January     April     July  October
## 1   6.666667 29.266667 55.06667 33.53333
## 2 -26.142857  3.571429 38.71429 10.28571
##
## Clustering vector:
##       Atlanta     Baltimore       Bismark        Boston       Chicago
##             1             1             2             1             2
##        Dallas        Denver       El Paso      Honolulu       Houston
##             1             2             1             1             1
##        Juneau   Los Angeles         Miami     Nashville      New York
##             2             1             1             1             1
##         Omaha       Phoenix      Portland          Reno San Francisco
##             2             1             2             2             1
##       Seattle    Washington
##             1             1
##
## Within cluster sum of squares by cluster:
## [1] 9572.933 1117.429
##  (between_SS / total_SS =  53.2 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

```r
lowtemp_km3 <- kmeans(lowtemp, 3)
lowtemp_km3
```

```
## K-means clustering with 3 clusters of sizes 9, 7, 6
##
## Cluster means:
##      January     April     July  October
## 1  -6.555556 22.555556 52.77778 27.33333
## 2 -26.142857  3.571429 38.71429 10.28571
## 3  26.500000 39.333333 58.50000 42.83333
##
## Clustering vector:
##       Atlanta     Baltimore       Bismark        Boston       Chicago
##             1             1             2             1             2
##        Dallas        Denver       El Paso      Honolulu       Houston
##             1             2             1             3             3
##        Juneau   Los Angeles         Miami     Nashville      New York
##             2             3             3             1             1
##         Omaha       Phoenix      Portland          Reno San Francisco
##             2             3             2             2             3
##       Seattle    Washington
##             1             1
##
## Within cluster sum of squares by cluster:
## [1]  758.000 1117.429 2885.167
##  (between_SS / total_SS =  79.2 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

```
par(mfrow = c(1,2))
plot(lowtemp_pc$scores[, 1:2], ylim = xlim, type = "n")
text(lowtemp_pc$scores[, 1:2], labels = as.numeric(lowtemp_km2$cluster), cex=0.8)

plot(lowtemp_pc$scores[, 1:2], ylim = xlim, type = "n")
text(lowtemp_pc$scores[, 1:2], labels = as.numeric(lowtemp_km3$cluster), cex=0.8)
```



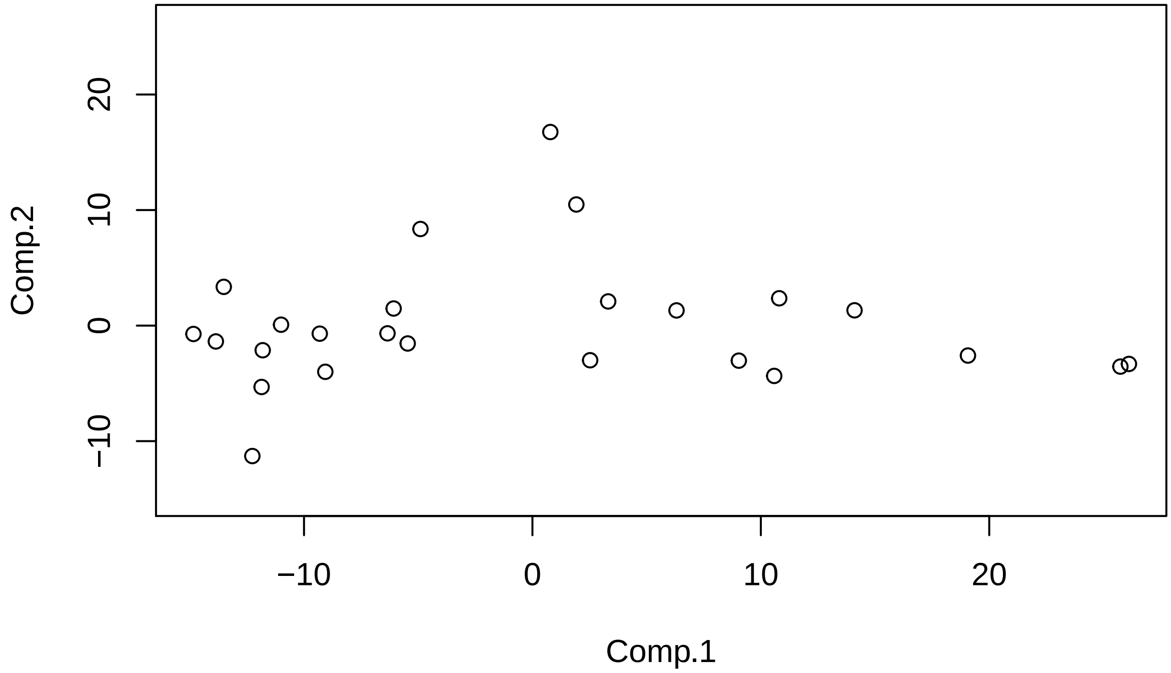
Try also other methods!

Exercise 17.5

```
protein <- read.table("data/protein.txt", sep = '\t', header = TRUE)
head(protein)
```

```
##                Rmeat Wmeat Eggs Milk Fish Cereals Sfoods Pulses Fruitveg
## Albania         10.1   1.4  0.5  8.9  0.2    42.3    0.6    5.5      1.7
## Austria          8.9  14.0  4.3 19.9  2.1    28.0    3.6    1.3      4.3
## Belgium         13.5   9.3  4.1 17.5  4.5    26.6    5.7    2.1      4.0
## Bulgaria         7.8   6.0  1.6  8.3  1.2    56.7    1.1    3.7      4.2
## Czechoslovakia   9.7  11.4  2.8 12.5  2.0    34.3    5.0    1.1      4.0
## Denmark         10.6  10.8  3.7 25.0  9.9    21.9    4.8    0.7      2.4
```

```
protein_pc <- princomp(protein)
xlim <- range(protein_pc$scores[, 1])
plot(protein_pc$scores[, 1:2], ylim = xlim)
```



```
# possibly 2 clusters?
```

Try some agglomerative methods and plot solutions in PC space.