Lab 3: NCI60 Data Example

Jonathan Bryan

August 31, 2018

```
library(ISLR)
nci.labs = NCI60$labs
nci.data = NCI60$data
dim(nci.data)
```

```
## [1]   64 6830
```

```
nci.labs[1:4]
```

```
## [1] "CNS"   "CNS"   "CNS"   "RENAL"
```

```
table(nci.labs)
```

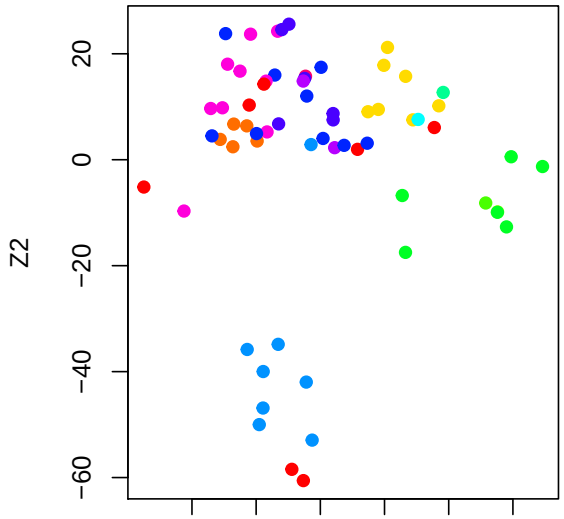
```
## nci.labs
##      BREAST         CNS       COLON K562A-repro K562B-repro    LEUKEMIA
##           7           5           7           1           1           6
## MCF7A-repro MCF7D-repro    MELANOMA       NSCLC     OVARIAN    PROSTATE
##           1           1           8           9           6           2
##       RENAL     UNKNOWN
##           9           1
```

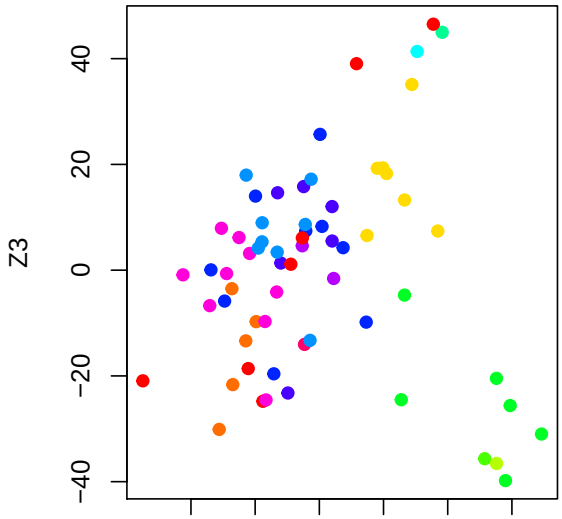
10.6.1 PCA on the NCI60 Data

```
#PCA
pr.out = prcomp(nci.data, scale = TRUE)

#assign colors to cell lines
Cols = function(vec){
  cols = rainbow(length(unique(vec)))
  return(cols[as.numeric(as.factor(vec))])
}
```

```
par(mfrow=c(1,2))
plot(pr.out$x[,1:2],
     col = Cols(nci.labs),
     pch = 19,
     xlab = "Z1",
     ylab = "Z2")
plot(pr.out$x[,c(1,3)],
     col = Cols(nci.labs),
     pch = 19,
     xlab = "Z1",
     ylab = "Z3")
```



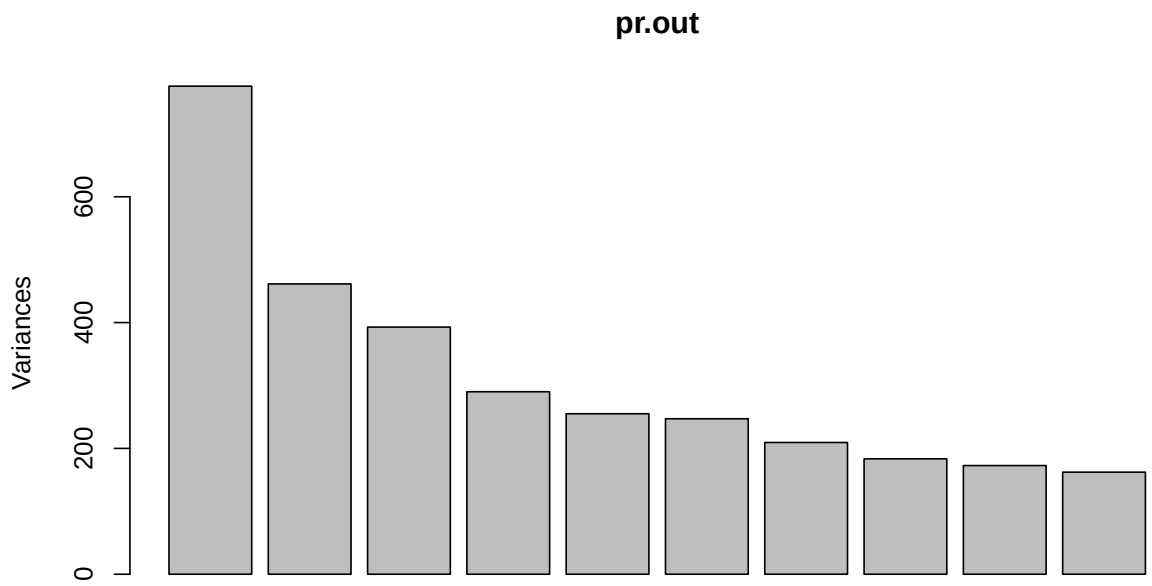



```
summary(pr.out)
```

```
## Importance of components:
##                             PC1      PC2      PC3      PC4      PC5
## Standard deviation     27.8535 21.48136 19.82046 17.03256 15.97181
## Proportion of Variance  0.1136  0.06756  0.05752  0.04248  0.03735
## Cumulative Proportion   0.1136  0.18115  0.23867  0.28115  0.31850
##                             PC6      PC7      PC8      PC9     PC10
## Standard deviation     15.72108 14.47145 13.54427 13.14400 12.73860
## Proportion of Variance  0.03619  0.03066  0.02686  0.02529  0.02376
## Cumulative Proportion   0.35468  0.38534  0.41220  0.43750  0.46126
##                            PC11     PC12     PC13     PC14     PC15
## Standard deviation     12.68672 12.15769 11.83019 11.62554 11.43779
## Proportion of Variance  0.02357  0.02164  0.02049  0.01979  0.01915
## Cumulative Proportion   0.48482  0.50646  0.52695  0.54674  0.56590
##                            PC16     PC17     PC18     PC19    PC20
## Standard deviation     11.00051 10.65666 10.48880 10.43518 10.3219
## Proportion of Variance  0.01772  0.01663  0.01611  0.01594  0.0156
## Cumulative Proportion   0.58361  0.60024  0.61635  0.63229  0.6479
##                            PC21    PC22    PC23    PC24    PC25    PC26
## Standard deviation     10.14608 10.0544 9.90265 9.64766 9.50764 9.33253
## Proportion of Variance  0.01507  0.0148 0.01436 0.01363 0.01324 0.01275
## Cumulative Proportion   0.66296  0.6778 0.69212 0.70575 0.71899 0.73174
##                            PC27    PC28    PC29    PC30    PC31    PC32
## Standard deviation     9.27320 9.0900 8.98117 8.75003 8.59962 8.44738
## Proportion of Variance 0.01259 0.0121 0.01181 0.01121 0.01083 0.01045
## Cumulative Proportion  0.74433 0.7564 0.76824 0.77945 0.79027 0.80072
##                            PC33    PC34    PC35    PC36    PC37    PC38
## Standard deviation     8.37305 8.21579 8.15731 7.97465 7.90446 7.82127
```

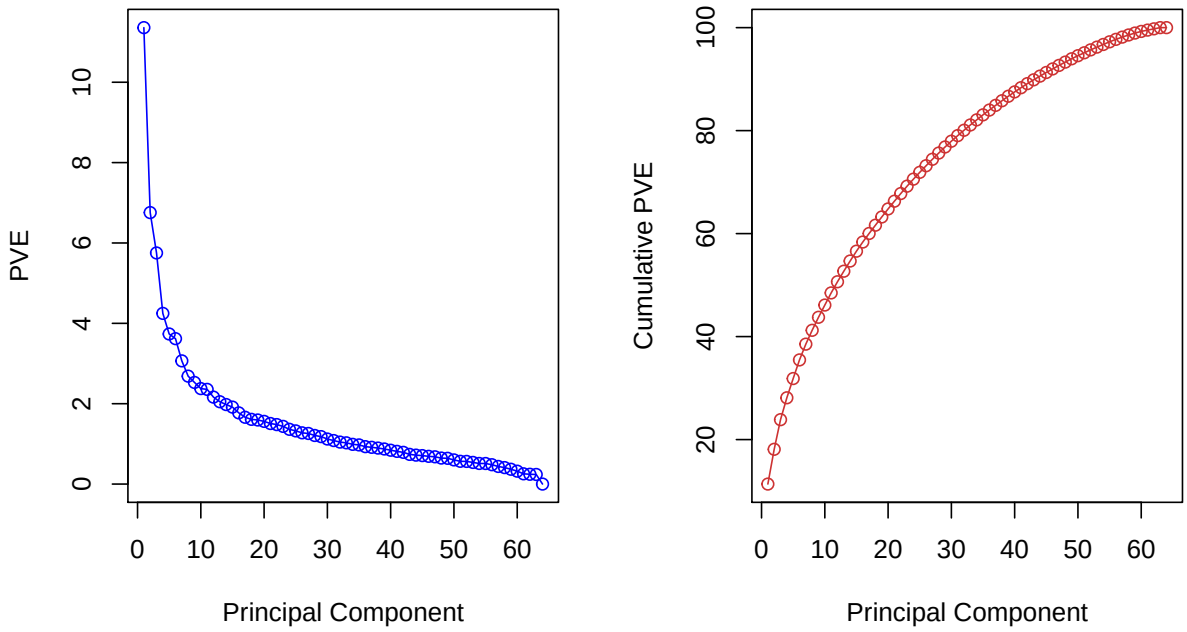
```
## Proportion of Variance 0.01026 0.00988 0.00974 0.00931 0.00915 0.00896
## Cumulative Proportion  0.81099 0.82087 0.83061 0.83992 0.84907 0.85803
##                           PC39    PC40    PC41   PC42    PC43   PC44
## Standard deviation     7.72156 7.58603 7.45619 7.3444 7.10449 7.0131
## Proportion of Variance 0.00873 0.00843 0.00814 0.0079 0.00739 0.0072
## Cumulative Proportion  0.86676 0.87518 0.88332 0.8912 0.89861 0.9058
##                           PC45   PC46    PC47    PC48    PC49    PC50
## Standard deviation     6.95839 6.8663 6.80744 6.64763 6.61607 6.40793
## Proportion of Variance 0.00709 0.0069 0.00678 0.00647 0.00641 0.00601
## Cumulative Proportion  0.91290 0.9198 0.92659 0.93306 0.93947 0.94548
##                           PC51    PC52    PC53    PC54    PC55    PC56
## Standard deviation     6.21984 6.20326 6.06706 5.91805 5.91233 5.73539
## Proportion of Variance 0.00566 0.00563 0.00539 0.00513 0.00512 0.00482
## Cumulative Proportion  0.95114 0.95678 0.96216 0.96729 0.97241 0.97723
##                           PC57   PC58    PC59    PC60    PC61    PC62
## Standard deviation     5.47261 5.2921 5.02117 4.68398 4.17567 4.08212
## Proportion of Variance 0.00438 0.0041 0.00369 0.00321 0.00255 0.00244
## Cumulative Proportion  0.98161 0.9857 0.98940 0.99262 0.99517 0.99761
##                           PC63      PC64
## Standard deviation     4.04124 2.148e-14
## Proportion of Variance 0.00239 0.000e+00
## Cumulative Proportion  1.00000 1.000e+00
```

```
#plot variance explained by first few principal components
plot(pr.out)
```




```
pve = 100*pr.out$sdev^2/sum(pr.out$sdev^2)
par(mfrow = c(1,2))
plot(pve,
     type = "o",
```

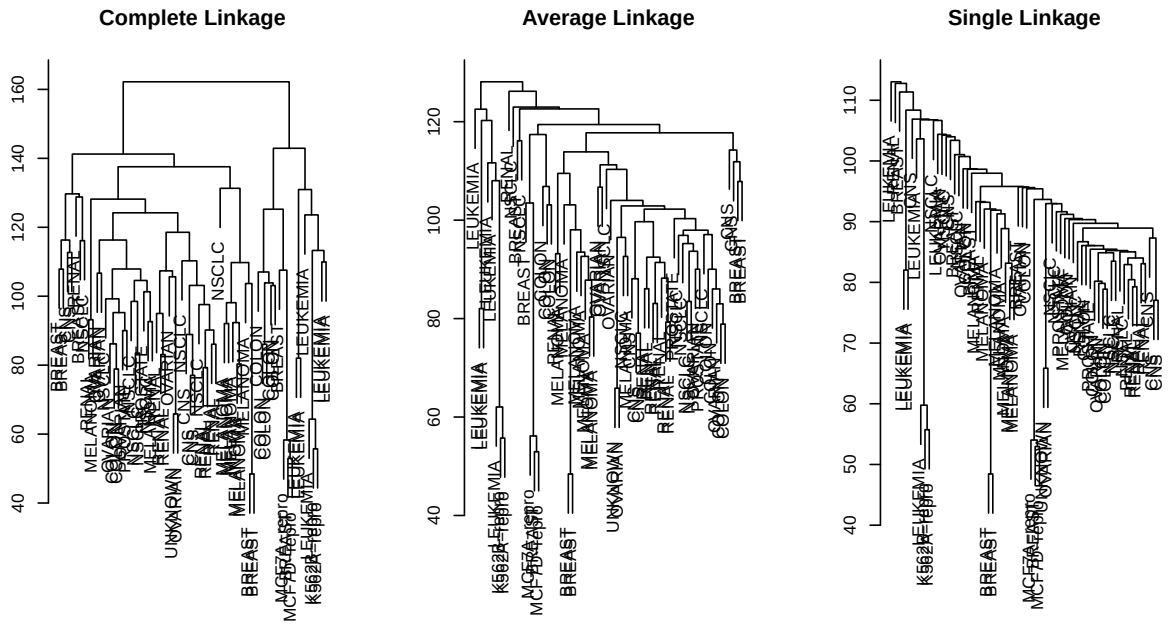
```
     ylab = "PVE",
     xlab = "Principal Component",
     col = "blue")
plot(cumsum(pve),
     type = "o",
     ylab = "Cumulative PVE",
     xlab = "Principal Component",
     col = "brown3")
```



10.6.2 Clustering the Observations of the NCI60 Data

```
sd.data = scale(nci.data)
par(mfrow = c(1,3))
data.dist = dist(sd.data)
plot(hclust(data.dist, method = "complete"),
     labels = nci.labs,
     main = "Complete Linkage",
     xlab = "",
     sub = "",
     ylab = "")
plot(hclust(data.dist, method = "average"),
     labels = nci.labs,
     main = "Average Linkage",
     xlab = "",
     sub = "",
     ylab = "")
plot(hclust(data.dist, method = "single"),
```

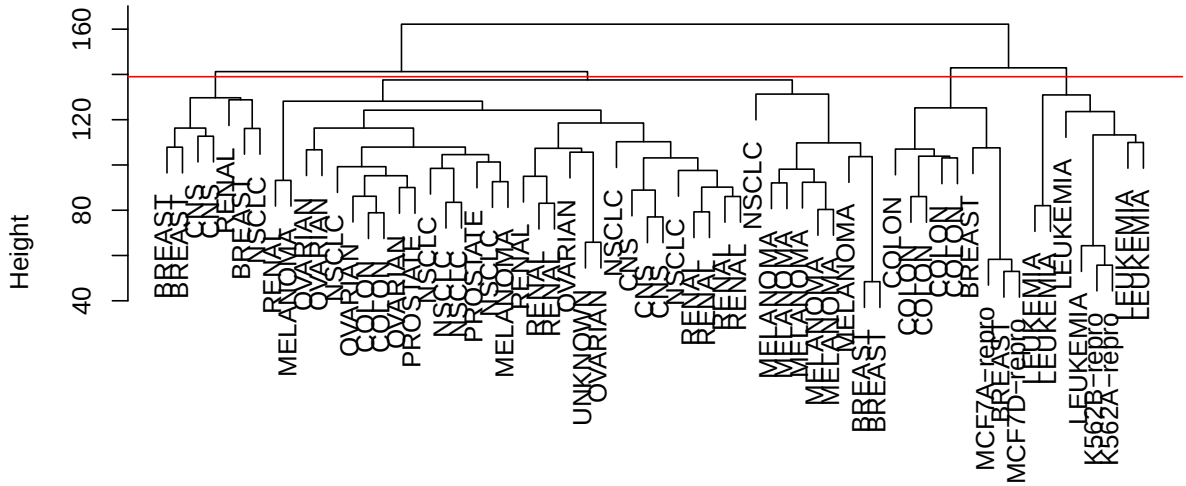
```
     labels = nci.labs,
     main = "Single Linkage",
     xlab = "",
     sub = "",
     ylab = "")
```




```
#proceed with complete linkage and cut at height to yield 4 clusters
hc.out = hclust(dist(sd.data))
hc.clusters = cutree(hc.out,4)
table(hc.clusters, nci.labs)
```

```
##             nci.labs
## hc.clusters BREAST CNS COLON K562A-repro K562B-repro LEUKEMIA MCF7A-repro
##            1      2   3     2           0           0        0           0
##            2      3   2     0           0           0        0           0
##            3      0   0     0           1           1        6           0
##            4      2   0     5           0           0        0           1
##             nci.labs
## hc.clusters MCF7D-repro MELANOMA NSCLC OVARIAN PROSTATE RENAL UNKNOWN
##            1           0        8     8       6        2     8       1
##            2           0        0     1       0        0     1       0
##            3           0        0     0       0        0     0       0
##            4           1        0     0       0        0     0       0
```

```
par(mfrow = c(1,1))
plot(hc.out, labels = nci.labs)
abline(h=139, col="red")
```



dist(sd.data)
hclust (*, "complete")

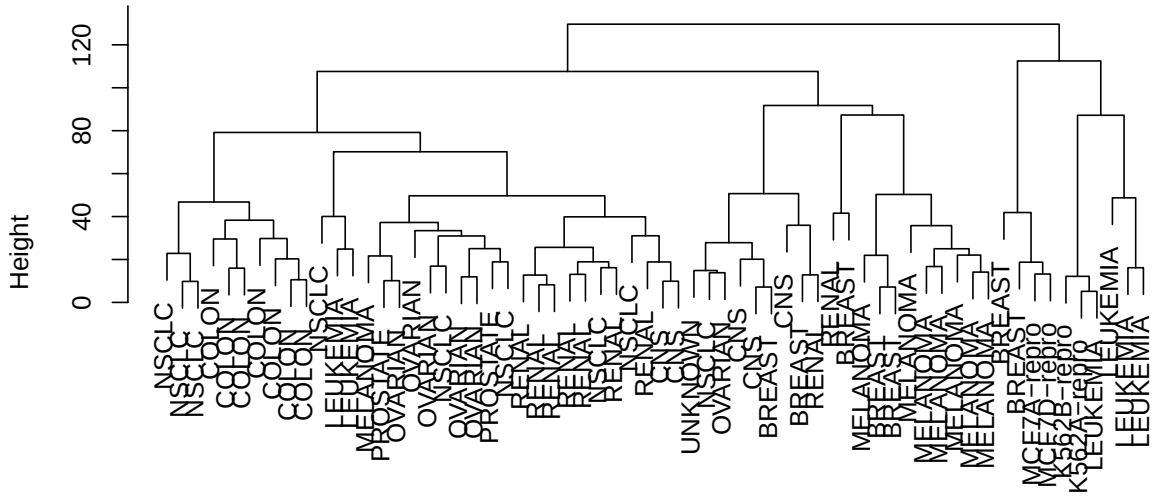
```
hc.out
```

```
##
## Call:
## hclust(d = dist(sd.data))
##
## Cluster method   : complete
## Distance         : euclidean
## Number of objects: 64
```

```
# k=4 clustering compared to hierarchical clustering
set.seed(2)
km.out = kmeans(sd.data, 4, nstart = 20)
km.clusters = km.out$cluster
table(km.clusters, hc.clusters)
```

```
##            hc.clusters
## km.clusters  1  2  3  4
##           1 11  0  0  9
##           2  0  0  8  0
##           3  9  0  0  0
##           4 20  7  0  0
```

```
#hierarchical clustering on first few principal compoenents
hc.out = hclust(dist(pr.out$x[,1:5]))
plot(hc.out, labels = nci.labs, main = "Hier. Clust. on First Five Score Vectors")
```




```
table(cutree(hc.out,4), nci.labs)
```

```
##    nci.labs
##     BREAST CNS COLON K562A-repro K562B-repro LEUKEMIA MCF7A-repro
##   1      0   2     7           0           0        2           0
##   2      5   3     0           0           0        0           0
##   3      0   0     0           1           1        4           0
##   4      2   0     0           0           0        0           1
##    nci.labs
##     MCF7D-repro MELANOMA NSCLC OVARIAN PROSTATE RENAL UNKNOWN
##   1           0        1     8       5        2     7       0
##   2           0        7     1       1        0     2       1
##   3           0        0     0       0        0     0       0
##   4           1        0     0       0        0     0       0
```