

Image Compression

Ngoc Duong

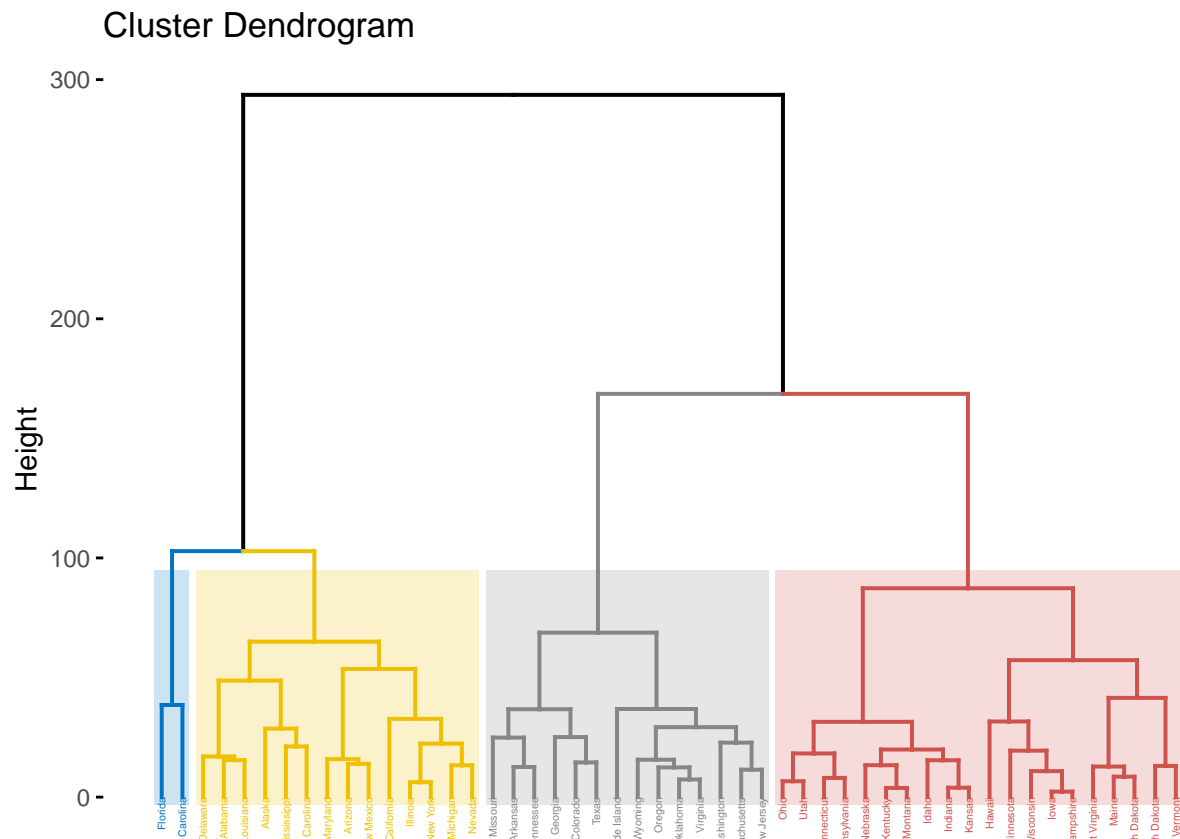
5/30/2020

Hierarchical Clustering

```
#load data  
data("USArrests")  
  
data = USArrests
```

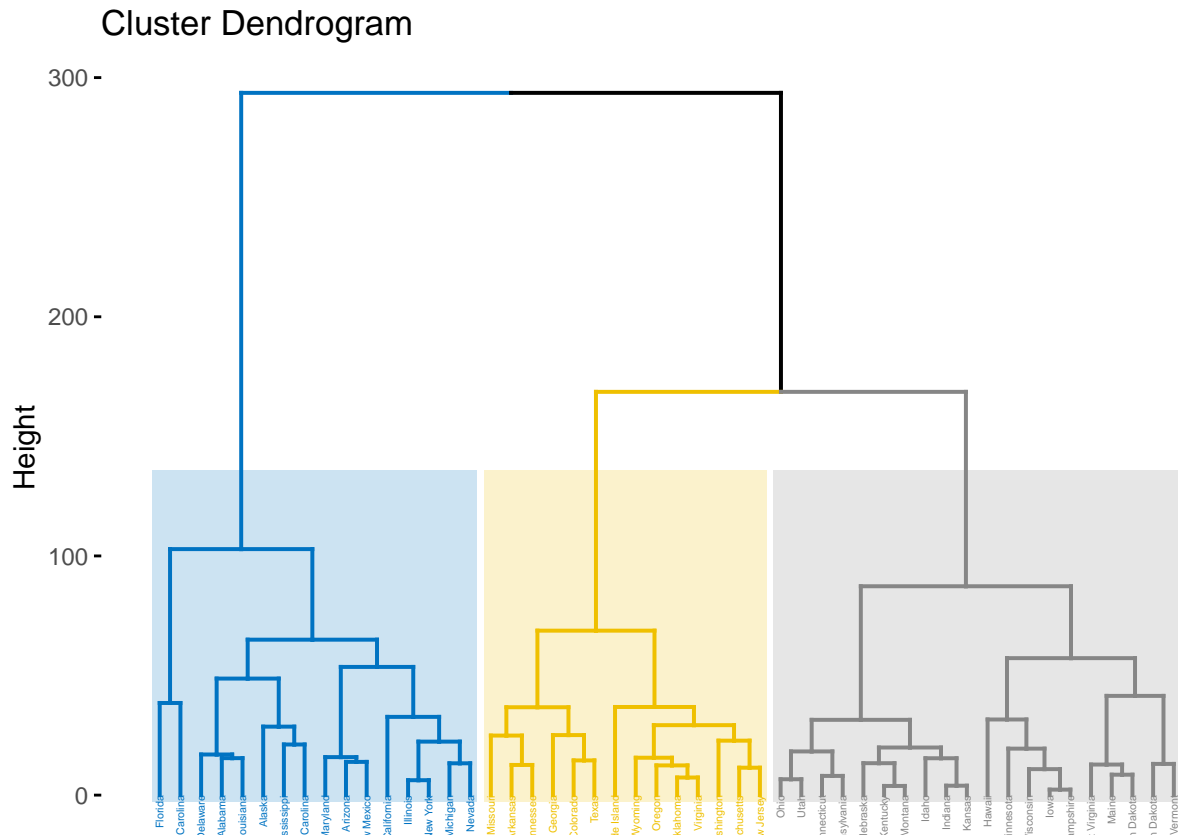
a) Cluster states using complete linkage and Euclidean distance

```
hc.complete <-hclust(dist(data), method = "complete")  
  
fviz_dend(hc.complete, k = 4,  
          cex = 0.3,  
          palette = "jco",  
          color_labels_by_k = TRUE,  
          rect = TRUE, rect_fill = TRUE,  
          rect_border = "jco", labels_track_height = 2.5)
```



b) Cut the dendrogram at height that results in 3 distinct clusters

```
fviz_dend(hc.complete, k = 3,
  cex = 0.3,
  palette = "jco",
  color_labels_by_k = TRUE,
  rect = TRUE, rect_fill = TRUE,
  rect_border = "jco", labels_track_height = 2.5)
```



Look at the states that are in each clusters

```
ind3.complete <-cutree(hc.complete, 3)
```

```
#cluster 1
```

```
data[ind3.complete==1,]
```

```
##           Murder  Assault  UrbanPop  Rape
## Alabama      13.2      236      58 21.2
## Alaska       10.0      263      48 44.5
## Arizona       8.1      294      80 31.0
## California    9.0      276      91 40.6
## Delaware      5.9      238      72 15.8
## Florida      15.4      335      80 31.9
## Illinois     10.4      249      83 24.0
## Louisiana    15.4      249      66 22.2
## Maryland     11.3      300      67 27.8
## Michigan     12.1      255      74 35.1
## Mississippi  16.1      259      44 17.1
## Nevada       12.2      252      81 46.0
## New Mexico   11.4      285      70 32.1
## New York     11.1      254      86 26.1
## North Carolina 13.0      337      45 16.1
## South Carolina 14.4      279      48 22.5
```

```
#cluster 2
```

```
data[ind3.complete==2,]
```

```
##           Murder  Assault  UrbanPop  Rape
```

```
## Arkansas      8.8      190      50 19.5
## Colorado      7.9      204      78 38.7
## Georgia       17.4     211      60 25.8
## Massachusetts 4.4      149      85 16.3
## Missouri      9.0      178      70 28.2
## New Jersey    7.4      159      89 18.8
## Oklahoma      6.6      151      68 20.0
## Oregon        4.9      159      67 29.3
## Rhode Island  3.4      174      87  8.3
## Tennessee    13.2     188      59 26.9
## Texas         12.7     201      80 25.5
## Virginia      8.5      156      63 20.7
## Washington    4.0      145      73 26.2
## Wyoming       6.8      161      60 15.6
```

```
#cluster 3
data[ind3.complete==3,]
```

```
##           Murder Assault UrbanPop Rape
## Connecticut 3.3      110      77 11.1
## Hawaii      5.3       46      83 20.2
## Idaho       2.6      120      54 14.2
## Indiana     7.2      113      65 21.0
## Iowa        2.2       56      57 11.3
## Kansas      6.0      115      66 18.0
## Kentucky    9.7      109      52 16.3
## Maine       2.1       83      51  7.8
## Minnesota   2.7       72      66 14.9
## Montana     6.0      109      53 16.4
## Nebraska    4.3      102      62 16.5
## New Hampshire 2.1      57      56  9.5
## North Dakota 0.8      45      44  7.3
## Ohio        7.3      120      75 21.4
## Pennsylvania 6.3      106      72 14.9
## South Dakota 3.8       86      45 12.8
## Utah        3.2      120      80 22.9
## Vermont     2.2       48      32 11.2
## West Virginia 5.7      81      39  9.3
## Wisconsin   2.6       53      66 10.8
```

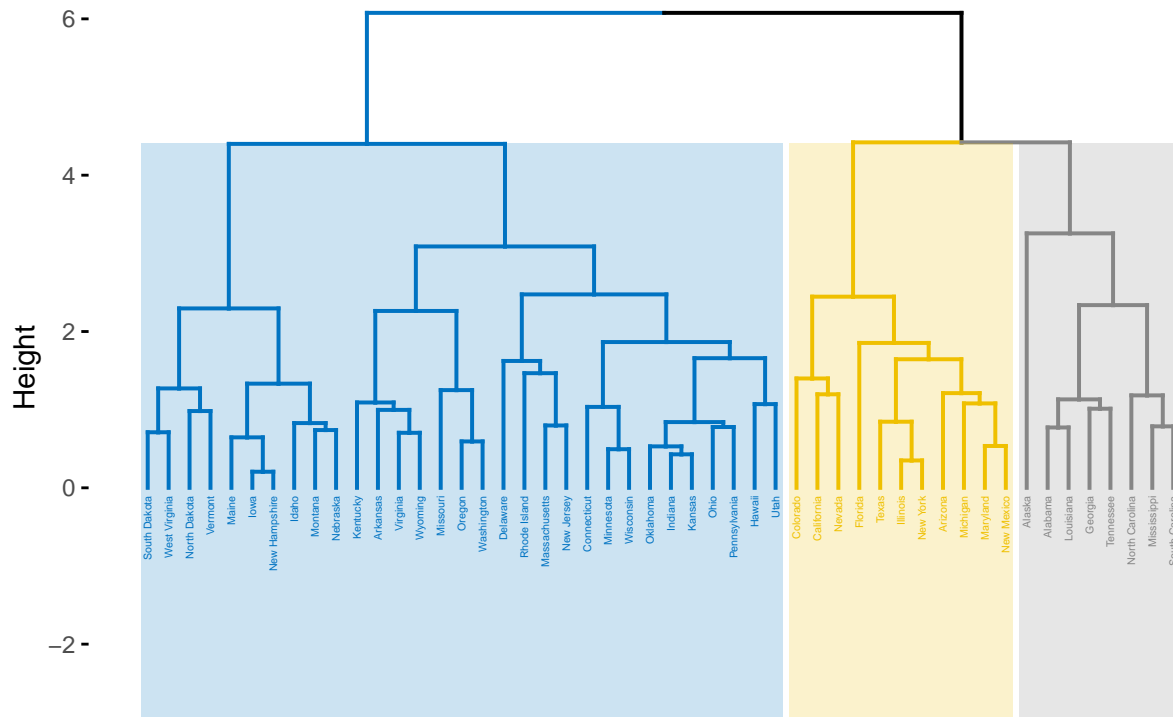
- c) Hierarchically cluster the states using complete linkage and Euclidean distance after scaling the variables to have $SD = 1$

```
dat1 <-scale(data)

hc.complete.scale <-hclust(dist(dat1), method = "complete")

fviz_dend(hc.complete.scale, k = 3,
  cex = 0.3,
  palette = "jco",
  color_labels_by_k = TRUE,
  rect = TRUE, rect_fill = TRUE,
  rect_border = "jco", labels_track_height = 2.5)
```

Cluster Dendrogram



```
ind3.complete.scale <-cutree(hc.complete.scale, 3)
```

```
#cluster 1
```

```
dat1[ind3.complete.scale==1,]
```

##		Murder	Assault	UrbanPop	Rape
##	Alabama	1.2425641	0.7828393	-0.52090661	-0.003416473
##	Alaska	0.5078625	1.1068225	-1.21176419	2.484202941
##	Georgia	2.2068599	0.4828549	-0.38273510	0.487701523
##	Louisiana	1.7476714	0.9388312	0.03177945	0.103348309
##	Mississippi	1.9083874	1.0588250	-1.48810723	-0.441152078
##	North Carolina	1.1966452	1.9947764	-1.41902147	-0.547916860
##	South Carolina	1.5180772	1.2988126	-1.21176419	0.135377743
##	Tennessee	1.2425641	0.2068693	-0.45182086	0.605142783

```
#cluster 2
```

```
dat1[ind3.complete.scale==2,]
```

##		Murder	Assault	UrbanPop	Rape
##	Arizona	0.07163341	1.4788032	0.9989801	1.0428784
##	California	0.27826823	1.2628144	1.7589234	2.0678203
##	Colorado	0.02571456	0.3988593	0.8608085	1.8649672
##	Florida	1.74767144	1.9707777	0.9989801	1.1389667
##	Illinois	0.59970018	0.9388312	1.2062373	0.2955249
##	Maryland	0.80633501	1.5507995	0.1008652	0.7012311
##	Michigan	0.99001041	1.0108275	0.5844655	1.4806140
##	Nevada	1.01296983	0.9748294	1.0680658	2.6443501
##	New Mexico	0.82929443	1.3708088	0.3081225	1.1603196
##	New York	0.76041616	0.9988281	1.4134946	0.5197310

```
## Texas      1.12776696 0.3628612 0.9989801 0.4556721
```

```
#cluster 3
```

```
dat1[ind3.complete.scale==3,]
```

##	Murder	Assault	UrbanPop	Rape
## Arkansas	0.23234938	0.23086801	-1.07359268	-0.18491660
## Connecticut	-1.03041900	-0.72908214	0.79172279	-1.08174077
## Delaware	-0.43347395	0.80683810	0.44629400	-0.57994629
## Hawaii	-0.57123050	-1.49704226	1.20623733	-0.11018125
## Idaho	-1.19113497	-0.60908837	-0.79724965	-0.75076995
## Indiana	-0.13500142	-0.69308401	-0.03730631	-0.02476943
## Iowa	-1.28297267	-1.37704849	-0.58999237	-1.06038781
## Kansas	-0.41051452	-0.66908525	0.03177945	-0.34506377
## Kentucky	0.43898421	-0.74108152	-0.93542116	-0.52656390
## Maine	-1.30593210	-1.05306531	-1.00450692	-1.43406455
## Massachusetts	-0.77786532	-0.26110644	1.34440885	-0.52656390
## Minnesota	-1.16817555	-1.18505846	0.03177945	-0.67603460
## Missouri	0.27826823	0.08687549	0.30812248	0.74393700
## Montana	-0.41051452	-0.74108152	-0.86633540	-0.51588743
## Nebraska	-0.80082475	-0.82507715	-0.24456358	-0.50521095
## New Hampshire	-1.30593210	-1.36504911	-0.65907813	-1.25256442
## New Jersey	-0.08908257	-0.14111267	1.62075188	-0.25965195
## North Dakota	-1.60440462	-1.50904164	-1.48810723	-1.48744694
## Ohio	-0.11204199	-0.60908837	0.65355127	0.01793648
## Oklahoma	-0.27275797	-0.23710769	0.16995096	-0.13153421
## Oregon	-0.66306820	-0.14111267	0.10086521	0.86137826
## Pennsylvania	-0.34163624	-0.77707965	0.44629400	-0.67603460
## Rhode Island	-1.00745957	0.03887798	1.48258036	-1.38068216
## South Dakota	-0.91562187	-1.01706718	-1.41902147	-0.90024064
## Utah	-1.05337842	-0.60908837	0.99898006	0.17808366
## Vermont	-1.28297267	-1.47304350	-2.31713632	-1.07106429
## Virginia	0.16347111	-0.17711080	-0.17547783	-0.05679886
## Washington	-0.86970302	-0.30910395	0.51537975	0.53040744
## West Virginia	-0.47939280	-1.07706407	-1.83353601	-1.27391738
## Wisconsin	-1.19113497	-1.41304662	0.03177945	-1.11377020
## Wyoming	-0.22683912	-0.11711392	-0.38273510	-0.60129925

- d) We observe that the cluster memberships changed after scaling the variables. The variables should be scaled before the inter-observation dissimilarities are computed in order to ensure equal weights given to every variable in **X** despite their different scales.

PCA

```
img <- readJPEG('piggy.jpg')
```

```
dim(img)
```

```
## [1] 233 233 3
```

```
r <- img[, ,1]
```

```
g <- img[, ,2]
```

```
b <- img[, ,3]
```

```
img.r.pca <- prcomp(r, center = FALSE)
```

```

img.g.pca <- prcomp(g, center = FALSE)
img.b.pca <- prcomp(b, center = FALSE)

rgb.pca <- list(img.r.pca, img.g.pca, img.b.pca)

# Approximate X with XV_kV_k^T
compress <- function(pr, k)
{
  compressed.img <- pr$x[,1:k] %*% t(pr$rotation[,1:k])
  compressed.img
}

# Using first 20 PCs
pca20 <- sapply(rgb.pca, compress, k = 20, simplify = "array")

writeJPEG(pca20, "pca20.jpeg")

# Try to increase the number of PCs!
pca50 <- sapply(rgb.pca, compress, k = 50, simplify = "array")

writeJPEG(pca50, "pca50.jpeg")

pca100 <- sapply(rgb.pca, compress, k = 100, simplify = "array")

writeJPEG(pca100, "pca100.jpeg")

pca200 <- sapply(rgb.pca, compress, k = 200, simplify = "array")

writeJPEG(pca200, "pca200.jpeg")

knitr::include_graphics("piggy.jpg")

```



Figure 1: Original photo

```
knitr::include_graphics("pca50.jpeg")
```



Figure 2: PCA50

```
knitr::include_graphics("pca100.jpeg")
```

```
knitr::include_graphics("pca200.jpeg")
```




Figure 3: PCA100



Figure 4: PCA200