Exploratory Data Mining via Search Strategies Lab #4

Ross Jacobucci & Kevin J. Grimm

Outline

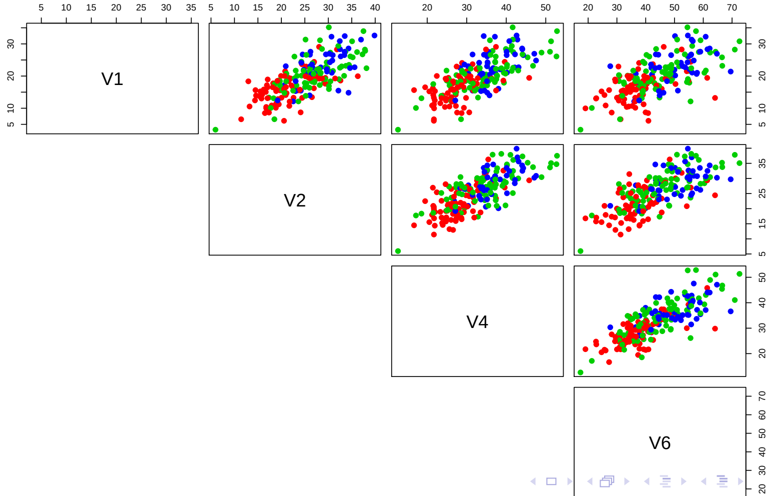
We will use some of the same packages used in the lectures to both clustering and finite mixture models.
To do this, we are going to use the WISC dataset that we will also use tomorrow. This is longitudinal data collected on kids in elementary school on verbal and performance scales along with mother's education.Data is WISC4VPE.DAT.

```
wisc <- read.table(
  "C:/Users/RJacobucci/Documents/GitHub/EDM_Labs/2015/wisc4vpe.dat")
names(wisc)<- c("V1","V2","V4","V6","P1","P2","P4", "P6", "Moeducat")
# note: V1 refers to verbal scores at grade 1, P is performance
```

Most analyses will not explicitly take into account the longitudinal nature of the data. However, we will look at a R package for longitudinal clustering at the end of the lab. In creating groups of individuals, we are going to compare these results to just classifying based on what their mother's education was.
First we will start with visualizing the data. Code adapted from : https://cran.r-project.org/web/packages/dendextend/vignettes/Cluster_Analysis.html

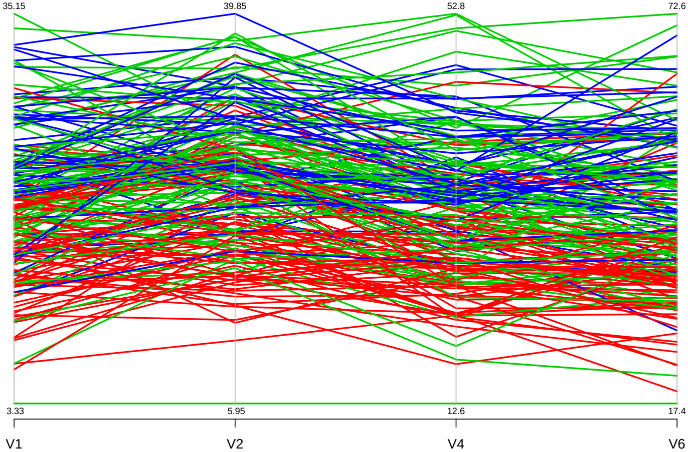
Visualize

```
col_class <- wisc$Moeducat + 2
# low = red, medium = green, high = blue
pairs(wisc[,1:4], col = col_class,lower.panel = NULL,
       cex.labels=2, pch=19, cex = 1.2)
```



Visualize

```r
MASS::parcoord(wisc[,1:4], col = col_class, var.label = TRUE, lwd = 2)
```

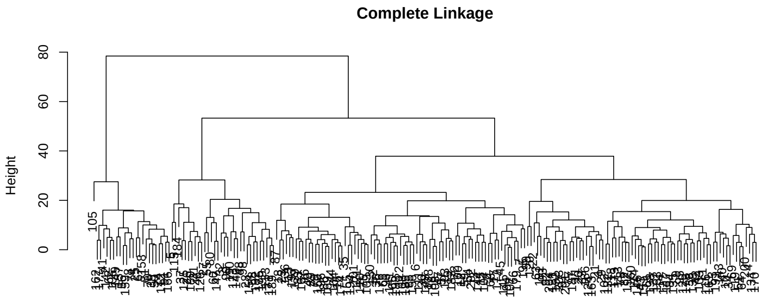


Hierarchical Clustering

Complete Linkage

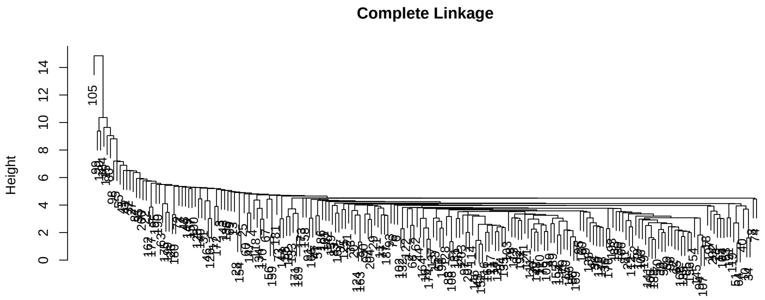
Using hclust() that is built into R

```r
wisc.dist <- dist(wisc[,1:4])
hc.clust.1 = hclust(wisc.dist, method='complete')
plot(hc.clust.1, main='Complete Linkage', xlab='', sub='', cex=.9)
```



Single Linkage

```r
wisc.dist <- dist(wisc[,1:4])
hc.clust.2 = hclust(wisc.dist, method='single')
plot(hc.clust.2, main='Complete Linkage', xlab='', sub='', cex=.9)
```



Hierarchical Clustering

Who is case #105?

```
wisc[105,1:4]
```

```
##            V1       V2       V4       V6
## 105 3.333333 5.952381 12.60417 17.35119
```

```
summary(wisc[,1:4])
```

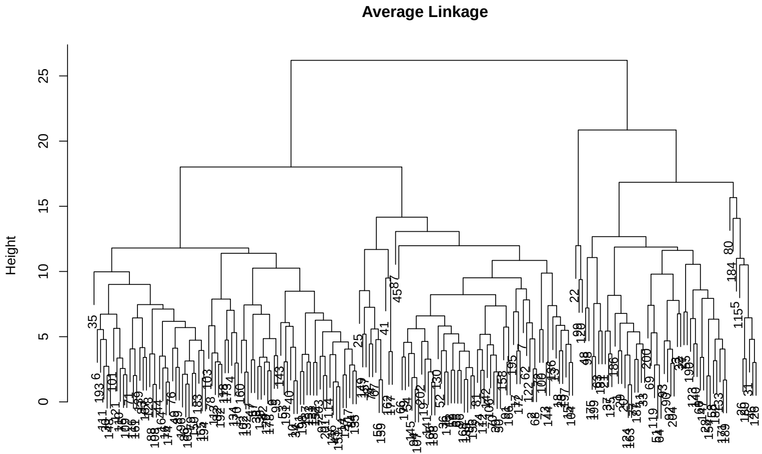
```
##        V1               V2               V4               V6
##  Min.   : 3.333   Min.   : 5.952   Min.   :12.60   Min.   :17.35
##  1st Qu.:15.636   1st Qu.:20.778   1st Qu.:27.24   1st Qu.:35.97
##  Median :19.330   Median :25.982   Median :32.82   Median :42.54
##  Mean   :19.585   Mean   :25.415   Mean   :32.61   Mean   :43.75
##  3rd Qu.:22.839   3rd Qu.:29.695   3rd Qu.:37.22   3rd Qu.:51.00
##  Max.   :35.149   Max.   :39.851   Max.   :52.84   Max.   :72.59
```

Hey, we found an outlier!

Average Linkage

```r
wisc.dist2 <- dist(wisc[-105,1:4])
hc.clust.3 = hclust(wisc.dist2, method='average')

plot(hc.clust.3, main='Average Linkage', xlab='', sub='', cex=.9)
```



Comparing Clusters to Mother's Education

Are we really just clustering people based on the family they come from?

```
pred.3 = cutree(hc.clust.3,3)
table(pred.3, wisc$Moeducat[-105]+1)
```

```
##
## pred.3  1  2  3
##      1  4 26 23
##      2 70 54 23
##      3  2  1  0
```

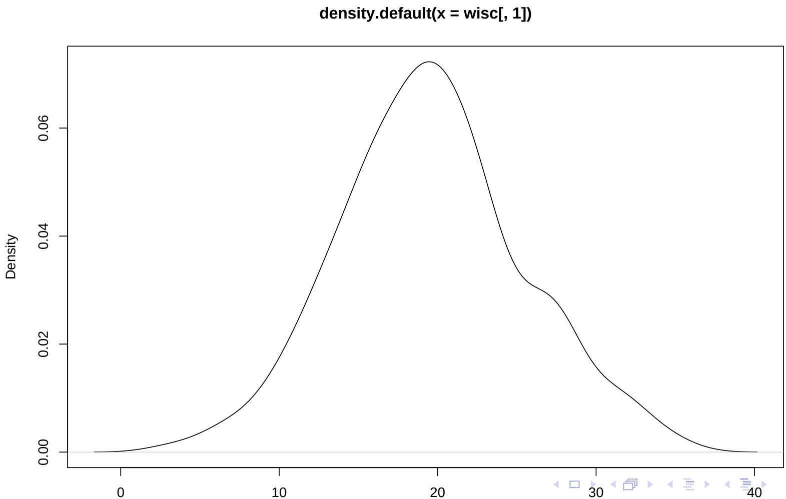
Looks like there is another factor involved other than mother's education

Finite Mixtures

Univariate Visualization

First Score

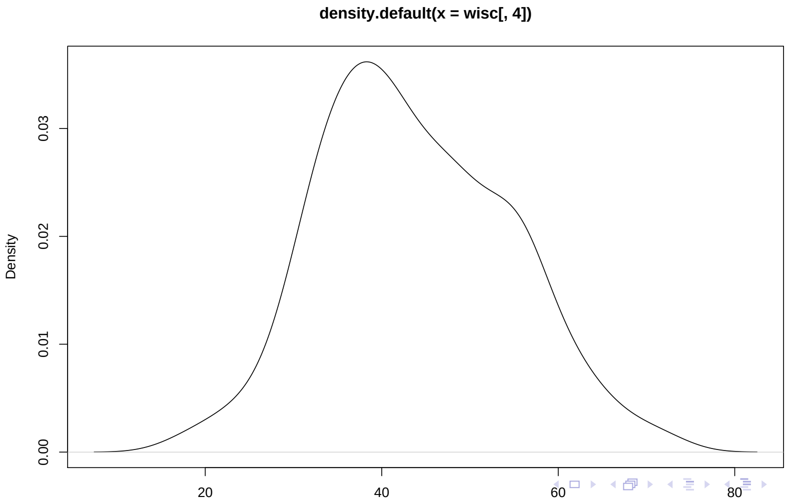
```
dd <- density(wisc[,1])
plot(dd)
```



Univariate Visualization

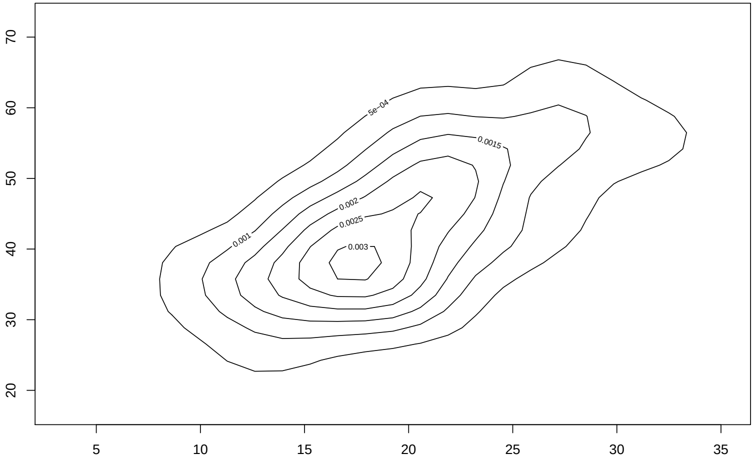
Last Score

```r
dd <- density(wisc[,4])
plot(dd)
```



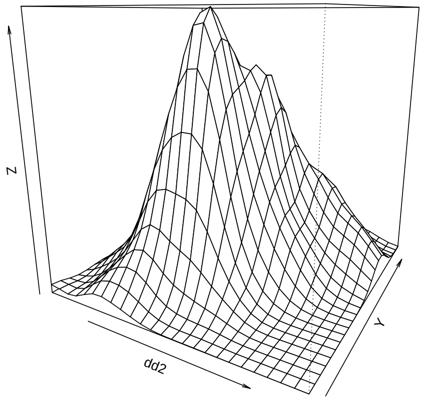
Bivariate Visualization

```
dd2 <- MASS::kde2d(wisc[,1],wisc[,4])
contour(dd2)
```



Bivariate Visualization

```
persp(dd2,theta=30,phi=15)
```



Mixture Visualization

What did we learn?

There seems to be some non-normality to the univariate and bivariate distributions. This means that finite mixtures are likely to find 2+ classes underlying the multivariate distribution