ETC3250 Lab 12 solution

Di Cook

Week 12

This lab is about diagnosing the results of cluster analysis. We will run different algorithms, and compare the results to determine which is better.

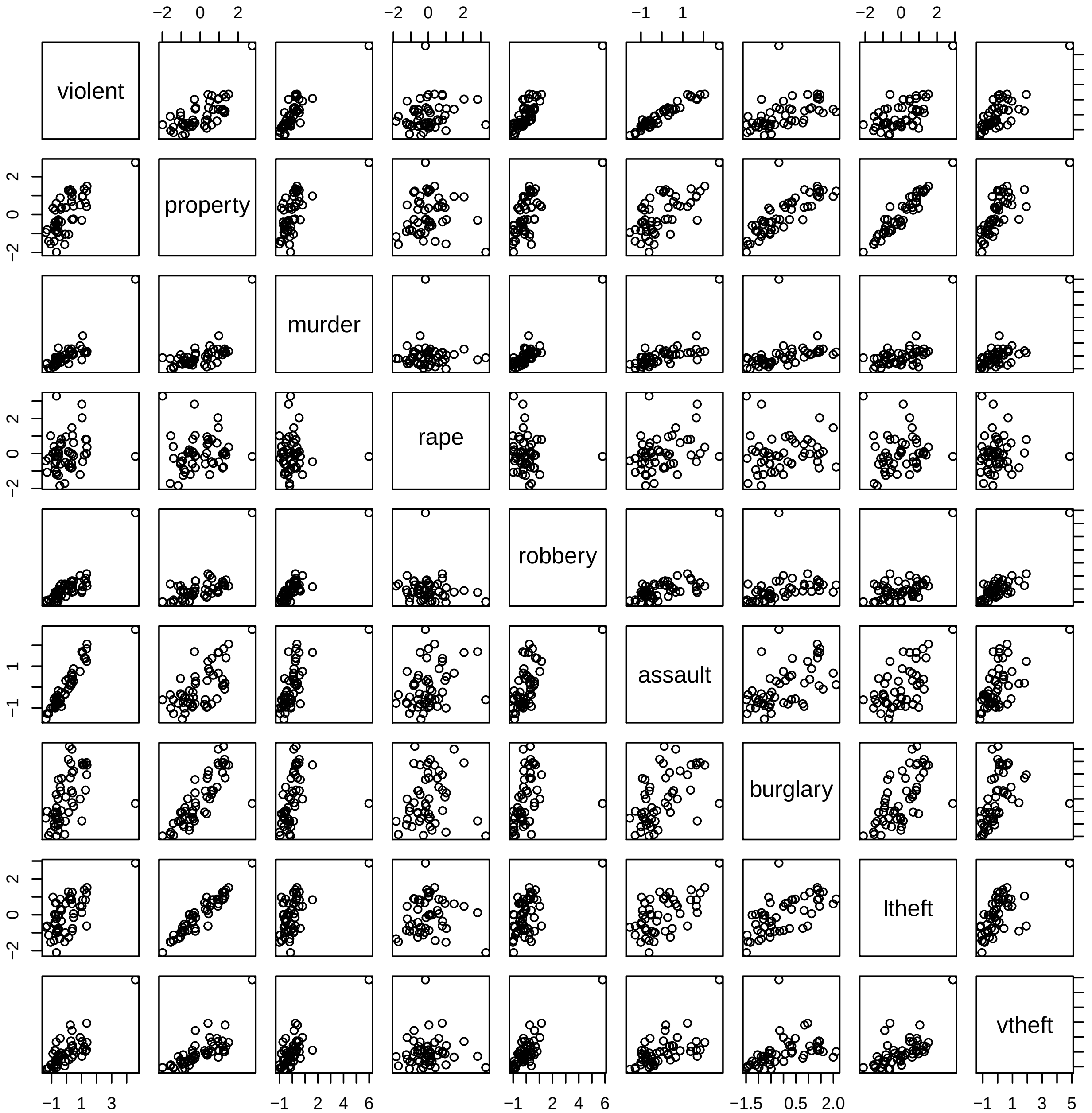
Data

The crime dataset (crimes.2008.csv) contains FBI crime rate statistics. These are the indices for 9 different types of crimes reported by the states of the USA, for 2008: violent, property, murder, rape, robbery, assault, burglary, ltheft (larceny theft), vtheft (vehicle theft). The values have been population adjusted so that the numers are per million people.

Question 1

Make a scatterplot matrix of the crime indices, with and without Washingto DC. Write a paragraph describing the relationships between the statistics, and about any observations about cluster patterns in the data.

```
pairs(crime[,2:10])
```

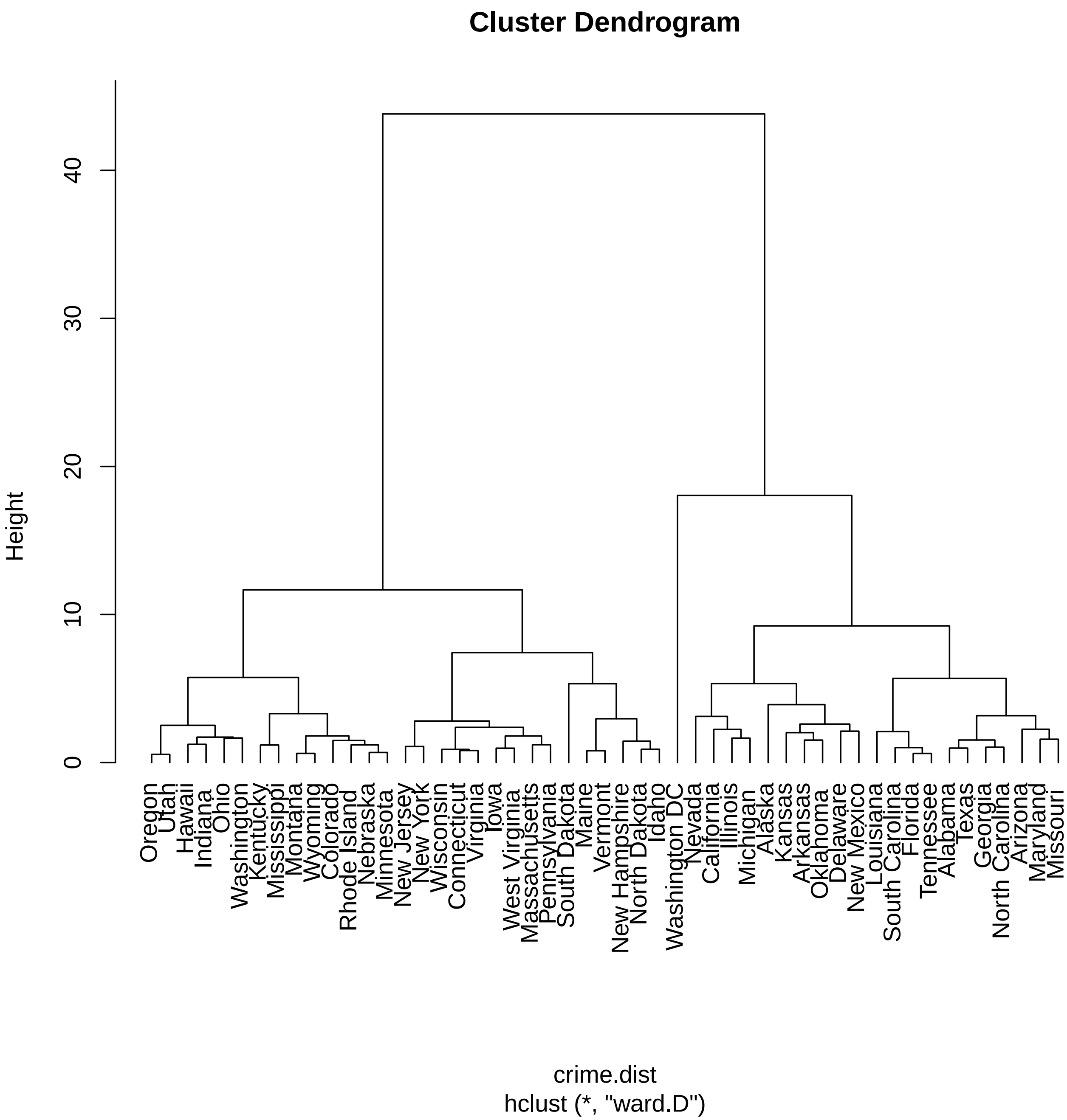


The pairwise relationships between the crime statistics reveal some outliers, and some positive association. One state has a very high rate of violent crimes, murder, robbery, larceny theft and vehicle theft. It also has the highest, albeit not by much, rate of property crime. (This is Washington, DC.) Removing this case makes it easier to read the associations. Most crime statistics show a positive association. The strongest relationships are between property and larceny theft, assault and violent crime. Rape has a different relationship with other crime rates. It has no association with murder, burglary and theft crimes, and a slightly negative association with robbery! There area few states that have high vehicle theft but relatively other types of crimes.

Question 2

Cluster the states using hierarchical clustering, with Euclidean distance and wards linkage. Plot the dendrogram. How many clusters would be suggested by the dendrogram?

```r
crime.dist <- dist(crime[,-1])
crime.hc <- hclust(crime.dist, method="ward.D")
plot(crime.hc, hang=-1)
```



2 or 3, mostly. It might be interesting to look at 4, 5, 6 or more clusters, too.

Question 3

Use k-means clustering with k set to several different values, say 2-8. Calculate the ratio of between Sum of Squares (SS) to total SS for each value of k. Tabulate this. What is between SS? total SS? What happens to this value as k ranges from 2 to 8? Why is this? Also, what happens if you change the random seed, which changes the initialization of k-means?

```
set.seed(407)
crime.km2 <- kmeans(crime[,-1], 2)
crime.km2$betweenss/crime.km2$totss
```

```
## [1] 0.390707
```

```
crime.km3 <- kmeans(crime[,-1], 3)
crime.km3$betweenss/crime.km3$totss
```

```
## [1] 0.6300192
```

```
crime.km4 <- kmeans(crime[,-1], 4)
crime.km4$betweenss/crime.km4$totss
```

```
## [1] 0.6809531
```

```
crime.km5 <- kmeans(crime[,-1], 5)
crime.km5$betweenss/crime.km5$totss
```

```
## [1] 0.744725
```

```
crime.km6 <- kmeans(crime[,-1], 6)
crime.km6$betweenss/crime.km6$totss
```

```
## [1] 0.7660463
```

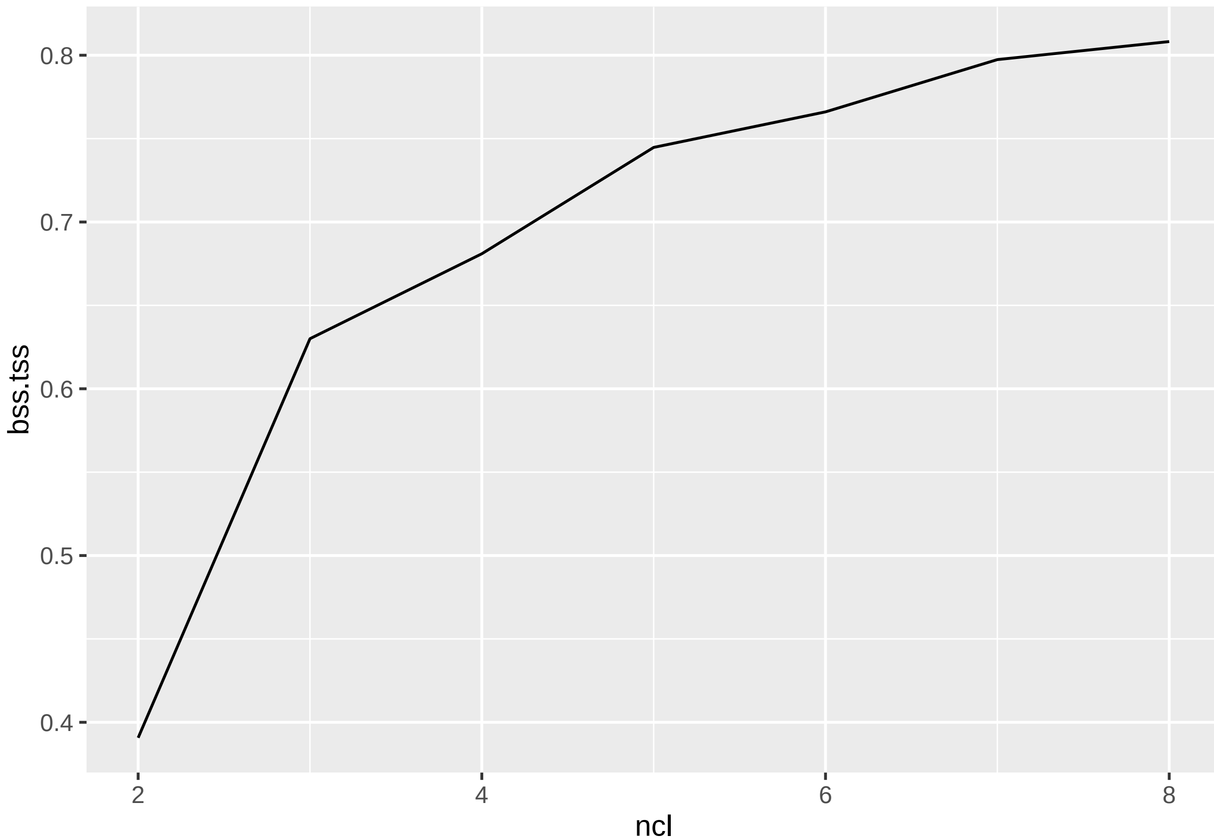
```
crime.km7 <- kmeans(crime[,-1], 7)
crime.km7$betweenss/crime.km7$totss
```

```
## [1] 0.7973879
```

```
crime.km8 <- kmeans(crime[,-1], 8)
crime.km8$betweenss/crime.km8$totss
```

```
## [1] 0.8081899
```

```
df <- data.frame(ncl=2:8, bss.tss = c(crime.km2$betweenss/crime.km2$totss, crime.km3$betweenss/crime.km3
qplot(ncl, bss.tss, data=df, geom="line")
```



It should increase. As more clusters are added the between cluster SS will be closer and closer to the total SS. Changing the initialization will change the results of the clustering.

Question 4

Use the *fpc* package in R, and the function *cluster.stats* to produce the statistic {`wb.ratio`} `to examine the within group distances to the between group distances for each cluster solution. How many clusters would be chosen by this approach? (The` *`wb.ratio`* `statistic reports the ratio between two quantities comparing within to between distances. The average of the distances between points that are in the same cluster, ie within. And the distances between points that are not in the same cluster, ie between. The smaller the value of this the better the result describes clustering as explaining the variation in the data.)`

```
cluster.stats(crime.dist, clustering=cutree(crime.hc, 3))$wb.ratio
```

```
## [1] 0.5237476
```

```
cluster.stats(crime.dist, clustering=cutree(crime.hc, 4))$wb.ratio
```

```
## [1] 0.5596829
```

```
cluster.stats(crime.dist, clustering=cutree(crime.hc, 5))$wb.ratio
```

```
## [1] 0.5194983
```

```
cluster.stats(crime.dist, clustering=cutree(crime.hc, 6))$wb.ratio
```

```
## [1] 0.5002843
```

```
cluster.stats(crime.dist, clustering=cutree(crime.hc, 7))$wb.ratio
```

```
## [1] 0.5039737
```

```
cluster.stats(crime.dist, clustering=cutree(crime.hc, 8))$wb.ratio
```

```
## [1] 0.4938994
```

```
cluster.stats(crime.dist, clustering=cutree(crime.hc, 9))$wb.ratio
```

```
## [1] 0.449124
```

```
cluster.stats(crime.dist, clustering=crime.km5$cluster)$wb.ratio
```

```
## [1] 0.4829496
```

The result using 3 clusters is better than 4, but 5, 6, 7, 8, 9 get sequentially lower values. 6, 7, 8 are all very similar so probably 5 is best from this group. The k-means with 5 clusters beats the hierarchical with 5 clusters.

Question 5

Decide on an appropriate number of clusters, and report the results. Tabulate the cluster means, standard deviation, and number of points in each cluster. Plot the cluster means using a parallel coordinate plot. List the states in each cluster. Write a paragraph describing the characteristics of each cluster, eg cluster 3 is characterized by low larceny and vehicle theft.

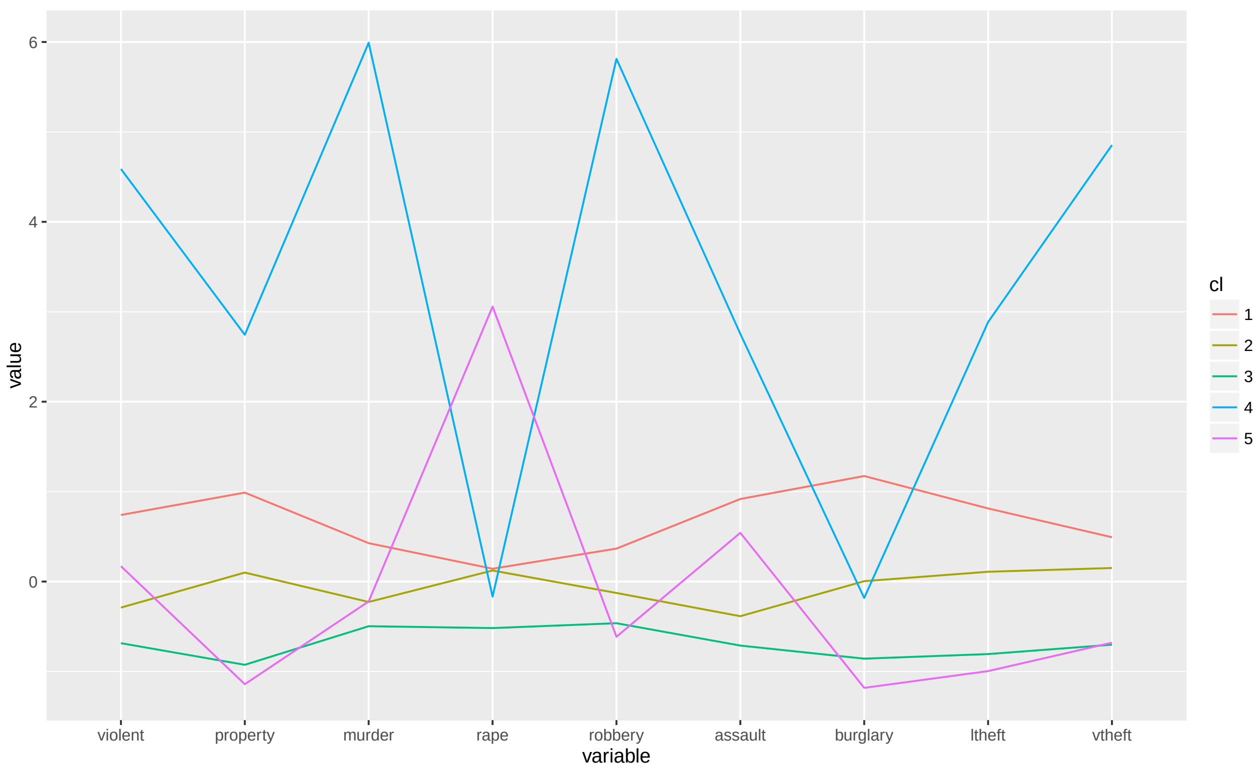
```
crime.km5$centers
```

```
##      violent    property     murder       rape    robbery    assault
## 1  0.7405402  0.98904408  0.4270204  0.1426699  0.3664575  0.9182316
## 2 -0.2893023  0.09990886 -0.2268656  0.1228799 -0.1268564 -0.3851498
## 3 -0.6851087 -0.92550594 -0.4964364 -0.5172863 -0.4631076 -0.7117160
## 4  4.5877209  2.74408978  5.9913967 -0.1671835  5.8122397  2.7542238
## 5  0.1708152 -1.14149865 -0.2210892  3.0577328 -0.6136905  0.5418111
##       burglary     ltheft     vtheft
## 1  1.173729481  0.8132411  0.4929179
## 2  0.004333686  0.1088870  0.1507216
## 3 -0.857350457 -0.8064014 -0.7021328
## 4 -0.182415612  2.8856250  4.8535616
## 5 -1.181967655 -0.9956927 -0.6795526
```

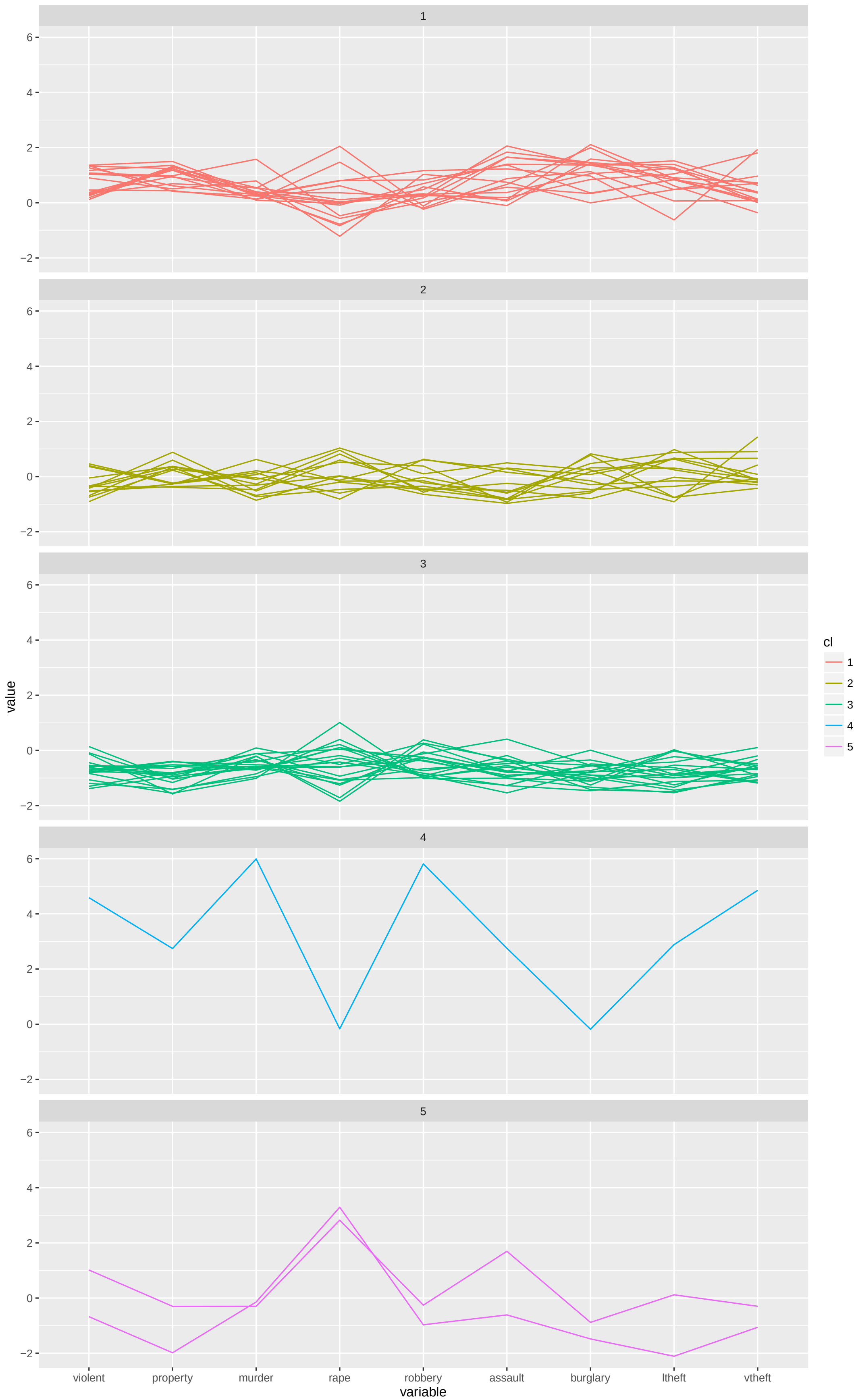
```r
# ddply(crime[,-1], .(crime.km3$cluster), colMeans)
# ddply(crime[,-1], .(crime.km3$cluster), function(x) sapply(x, sd))
crime.km.centers <- ddply(crime[,-1], .(crime.km5$cluster), colMeans)
crime.km.centers
```

```
##   crime.km5$cluster    violent    property     murder       rape
## 1                 1  0.7405402  0.98904408  0.4270204  0.1426699
## 2                 2 -0.2893023  0.09990886 -0.2268656  0.1228799
## 3                 3 -0.6851087 -0.92550594 -0.4964364 -0.5172863
## 4                 4  4.5877209  2.74408978  5.9913967 -0.1671835
## 5                 5  0.1708152 -1.14149865 -0.2210892  3.0577328
##      robbery    assault     burglary     ltheft     vtheft
## 1  0.3664575  0.9182316  1.173729481  0.8132411  0.4929179
## 2 -0.1268564 -0.3851498  0.004333686  0.1088870  0.1507216
## 3 -0.4631076 -0.7117160 -0.857350457 -0.8064014 -0.7021328
## 4  5.8122397  2.7542238 -0.182415612  2.8856250  4.8535616
## 5 -0.6136905  0.5418111 -1.181967655 -0.9956927 -0.6795526
```

```r
colnames(crime.km.centers)[1] <- "cl"
crime.km.centers$cl <- factor(crime.km.centers$cl)
ggparcoord(crime.km.centers, columns=2:10, groupColumn=1, scale="globalminmax")
```




```r
crime$cl <- crime.km5$cluster
crime$cl <- factor(crime$cl)
crime.m <- melt(crime, id.vars=c("State","cl"))
ggplot(crime.m, aes(x=variable, y=value, group=State, colour=cl)) + geom_line() + facet_wrap(~cl, ncol=
```


1
2
3
4
5
6
4
2
0
−2
value
variable
violent
property
murder
rape
robbery
assault
burglary
ltheft
vtheft
cl
1
2
3
4
5

```r
crime[crime$cl==1,1]
```

```
##  [1] Alabama        Arkansas       Arizona        Delaware
##  [5] Florida        Georgia        North Carolina Louisiana
##  [9] Maryland       Missouri       New Mexico     Nevada
## [13] Oklahoma       South Carolina Tennessee      Texas
## 51 Levels: Alabama Alaska Arizona Arkansas California ... Wyoming
```

```r
crime[crime$cl==2,1]
```

```
##  [1] California  Colorado    Hawaii      Nebraska    Illinois
##  [6] Indiana     Kansas      Michigan    Mississippi Ohio
## [11] Oregon      Utah        Washington
## 51 Levels: Alabama Alaska Arizona Arkansas California ... Wyoming
```

```r
crime[crime$cl==3,1]
```

```
##  [1] Connecticut    Iowa           Montana        North Dakota
##  [5] New Hampshire  New Jersey     Idaho          Kentucky
##  [9] Massachusetts  Maine          Minnesota      New York
## [13] Pennsylvania   Rhode Island   Virginia       Vermont
## [17] Wisconsin      West Virginia  Wyoming
## 51 Levels: Alabama Alaska Arizona Arkansas California ... Wyoming
```

```r
crime[crime$cl==4,1]
```

```
## [1] Washington DC
## 51 Levels: Alabama Alaska Arizona Arkansas California ... Wyoming
```

```r
crime[crime$cl==5,1]
```

```
## [1] South Dakota Alaska
## 51 Levels: Alabama Alaska Arizona Arkansas California ... Wyoming
```

Five clusters is really enough to summarize the cities. If you look at the 7 cluster solution it is hard to characterize all the clusters as different from each other. Clusters 1, 2, 3 are generally consistent across all variables, and are lowest, medium, highest crime, respectively. Cluster 4 is Washingto DC, and it has high crime on all factors except rape and burglary. Cluster 5 (Alaska and South Dakota) is distinguished by having abnormally high rape statistics. The clusters we get reflects, to a large extent that we used overall counts, so large states will appear together, and small states together. If we had first calculated crime per 1000 people, the results would change.