Cluster Analysis and other unsupervised learning methods

Nonparametric density clustering, Part 2

Werner Stuetzle
Department of Statistics
University of Washington

March 28, 2017

```
## Preamble

stat593.dir <- "/Users/wxs/Dropbox/Unsupervised-learning-spring-2016"
dir.sep <- "/"

data.dir <- paste(stat593.dir, "Data", sep = dir.sep)
tc.dir <- paste(data.dir, "Test-collection", sep = dir.sep)

echo <- F
## quartz()

source(paste(stat593.dir, "Code", "gsl-functions-5-28-2010.R",
             sep = dir.sep), echo = echo)

source(paste(data.dir, "generate-artificial-data-functions-3-19-2014.R",
             sep = dir.sep), echo = echo)

source(paste(stat593.dir, "A7-GKL-plots-assessing-cluster-separation",
             "assessment-pruning-functions-2-15-2017.R",
             sep = dir.sep))

library(MASS)
library(cluster)
library(colorspace)
library(mvtnorm)
```

```
## library(mclust)

options(expressions = 10000)

opts_chunk$set(fig.width=5, fig.height=5)
```

Runt pruning

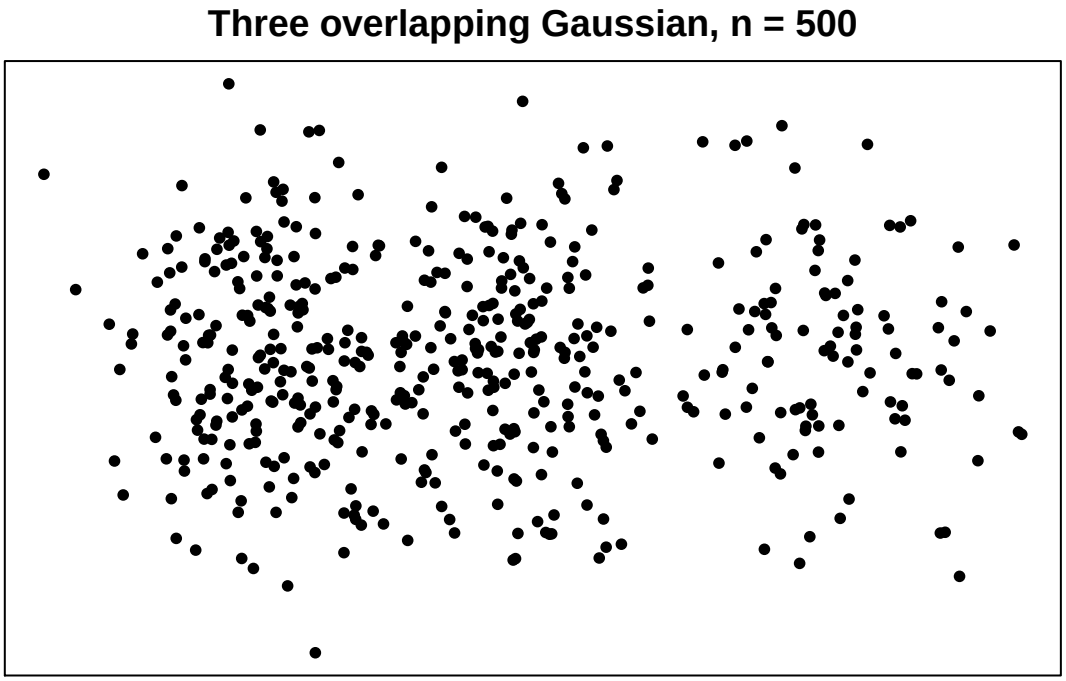
Runt size of a dendogram node = minimum of the number of leaves (observations) of the two subtrees rooted at the node.

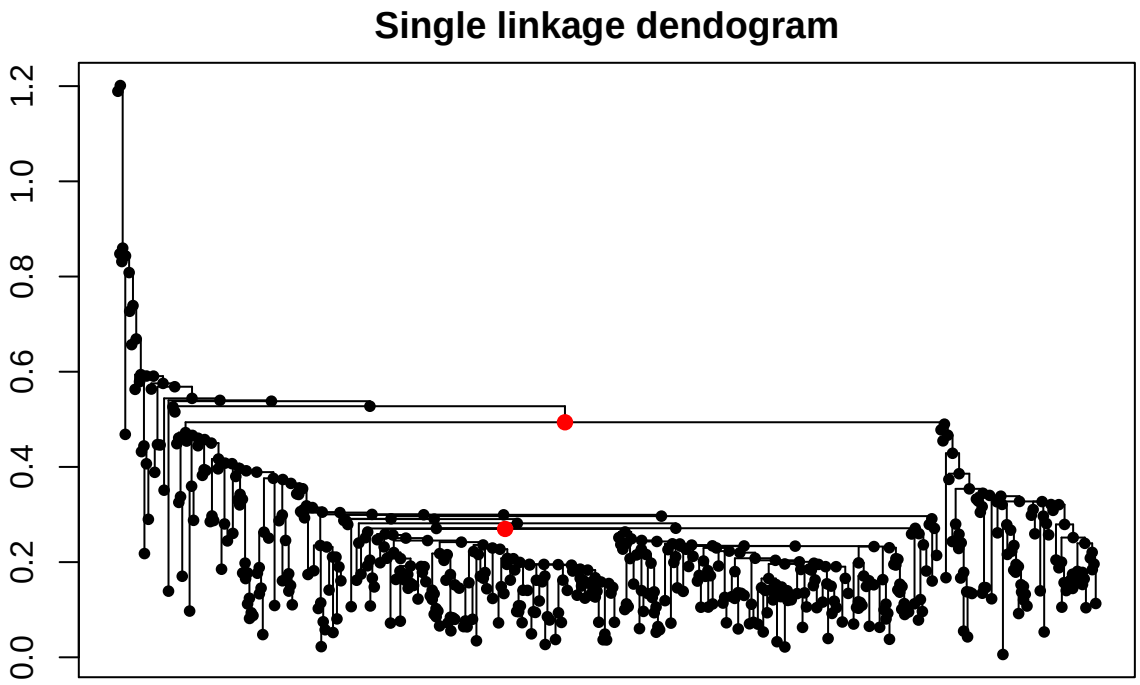
Single linkage merge process starts, and fragments grow, in high density regions (around modes)

Eventually the fragments growing around the various modes will be joined ⇒ dendogram nodes rooting two large subtrees ⇒ dendogram nodes with large runt size.

Dendogram nodes with large runt size indicate multimodality

Illustration

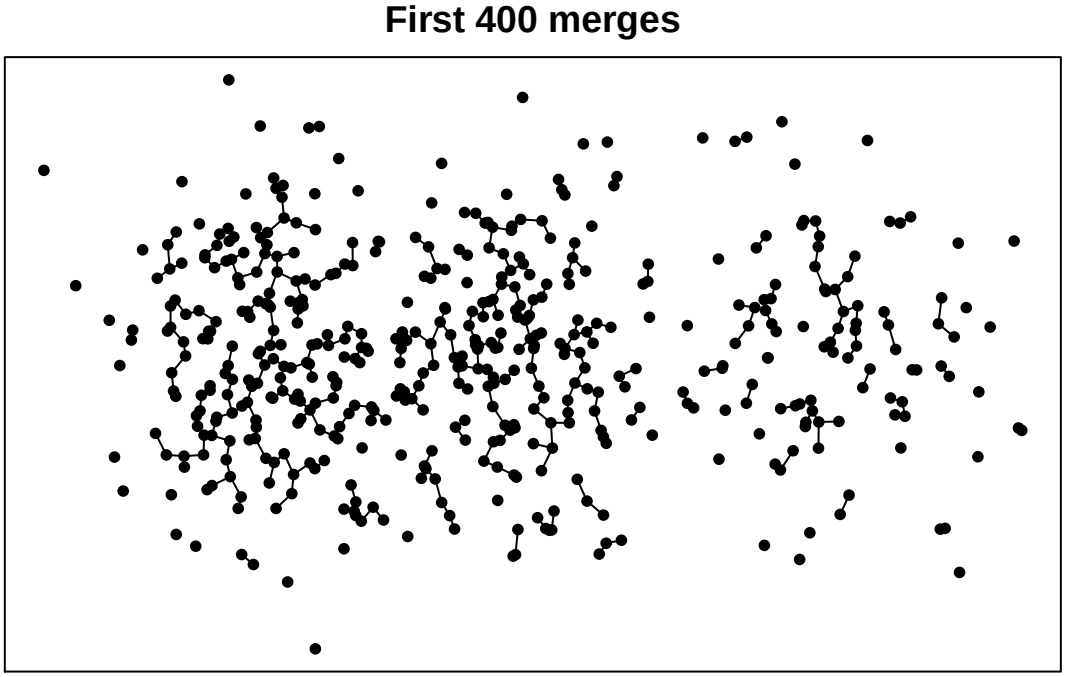




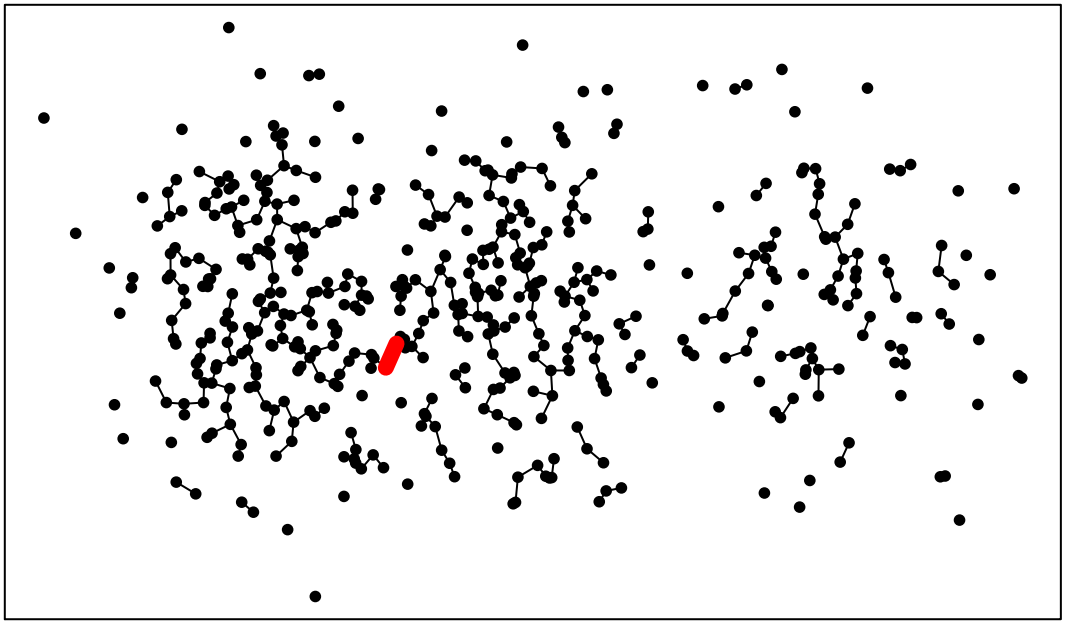
Dendogram cutting fails badly

Largest runt sizes are $123, 81, 30, 22, 22, 21, 17, 16, 14, 13, \ldots$.

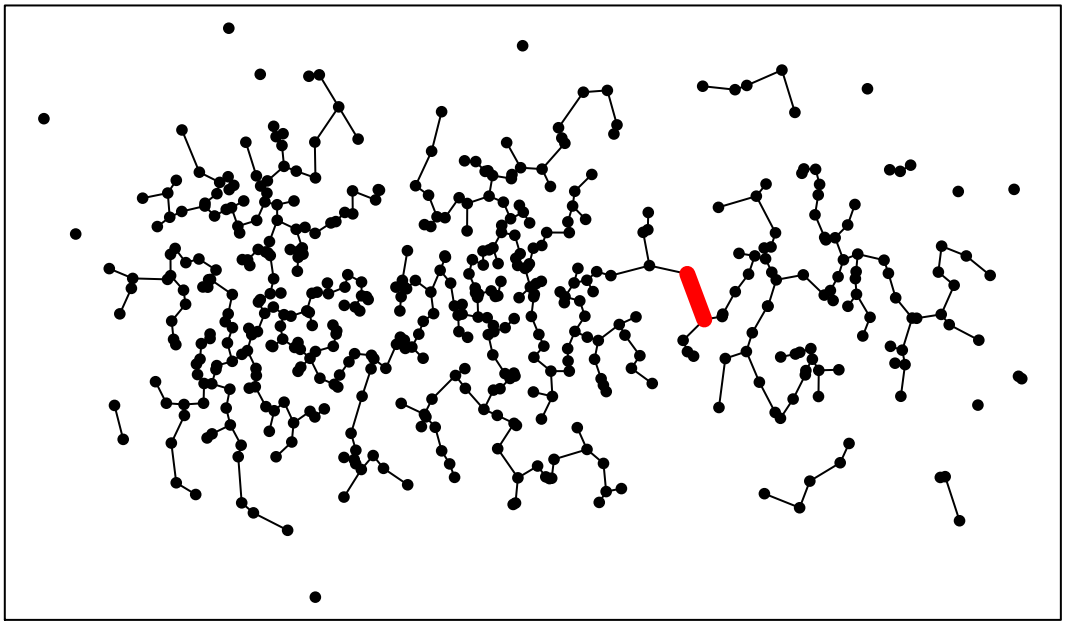
Dendogram nodes with two largest runt sizes are drawn in red.

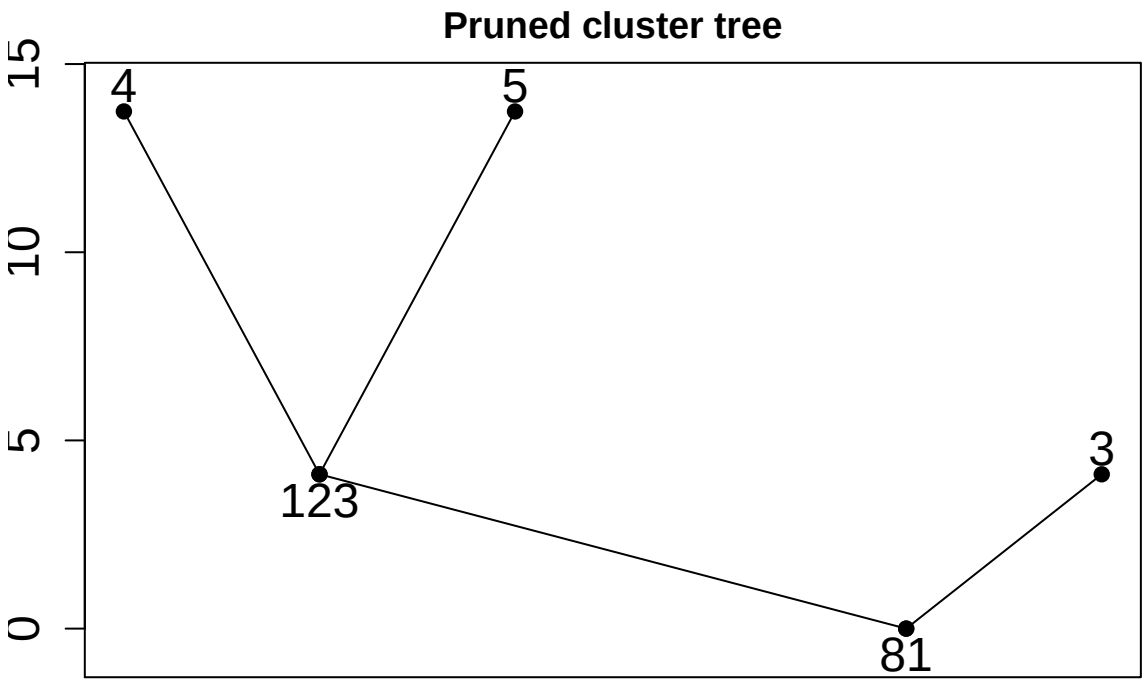


First merge of large fragments; runt size = 138



Second merge of large fragments; runt size = 88






```
## [1] "Table of group.id vs cluster.id"
##          cluster.id
## group.id   3   4   5
##        1   0 191   9
##        2   6  18 176
##        3  96   0   4
```

Single linkage with runt pruning does an almost perfect job