CS282R - Sepsis Cluster Optimzation

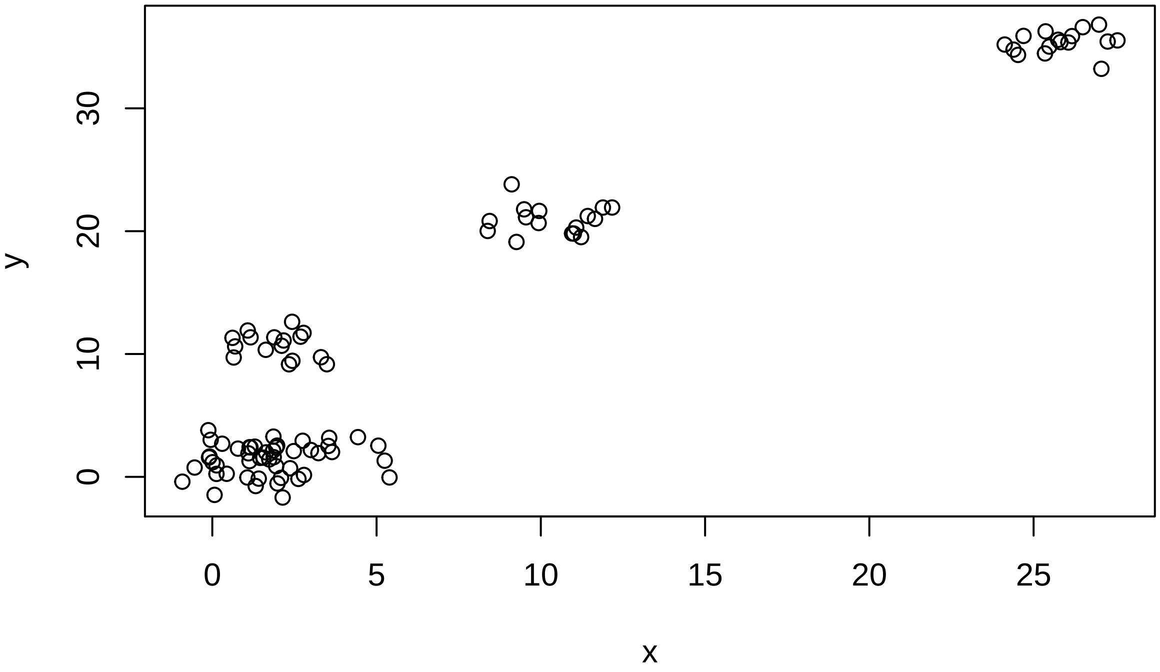
David Wihl

11/3/2017

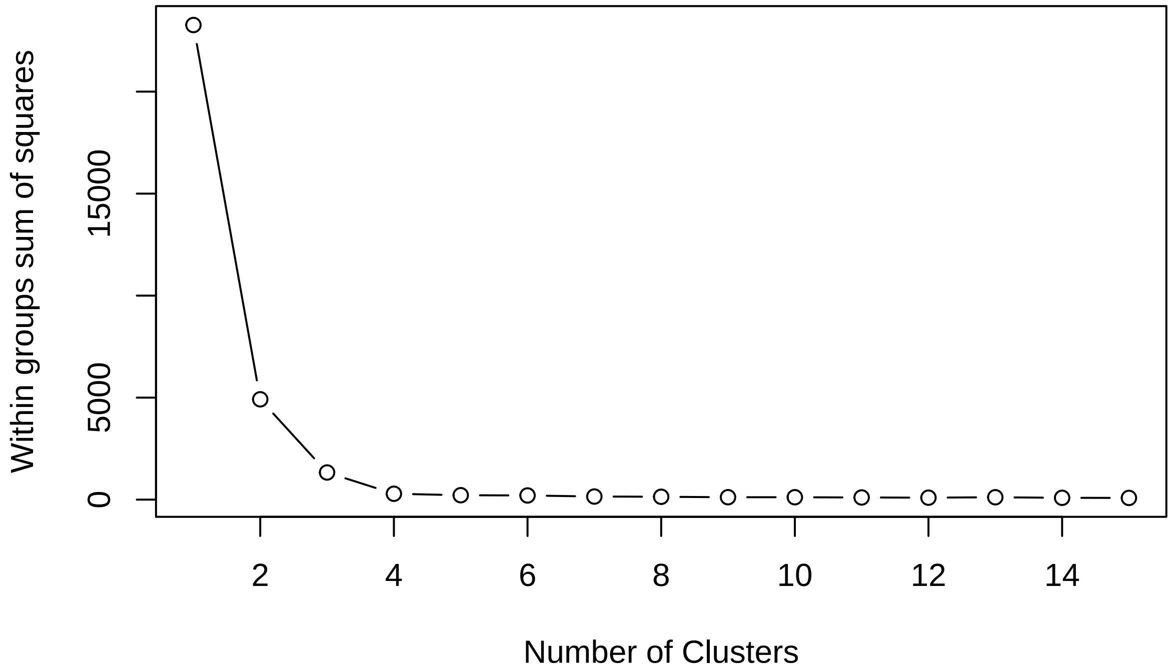
Using the suggestions from this StackOverflow question, this note will attempt to find the optimal clustering for the Sepsis data set.

```r
lots_of_clusters = function(d, in_k) {
  mydata <- d
  wss <- (nrow(mydata) - 1) * sum(apply(mydata, 2, var))
    for (i in 2:15)
      wss[i] <- sum(kmeans(mydata,centers = i)$withinss)
  plot(1:15,wss,type = "b",xlab = "Number of Clusters",ylab = "Within groups sum of squares")
  pamk.best <- pamk(d)
  cat("number of clusters estimated by optimum average silhouette width:",pamk.best$nc,"\n")
  plot(pam(d, pamk.best$nc))
  asw <- numeric(20)
  for (k in 2:20)
    asw[[k]] <- pam(d, k)$silinfo$avg.width
  k.best <- which.max(asw)
  cat("silhouette-optimal number of clusters:", k.best, "\n")
}
```

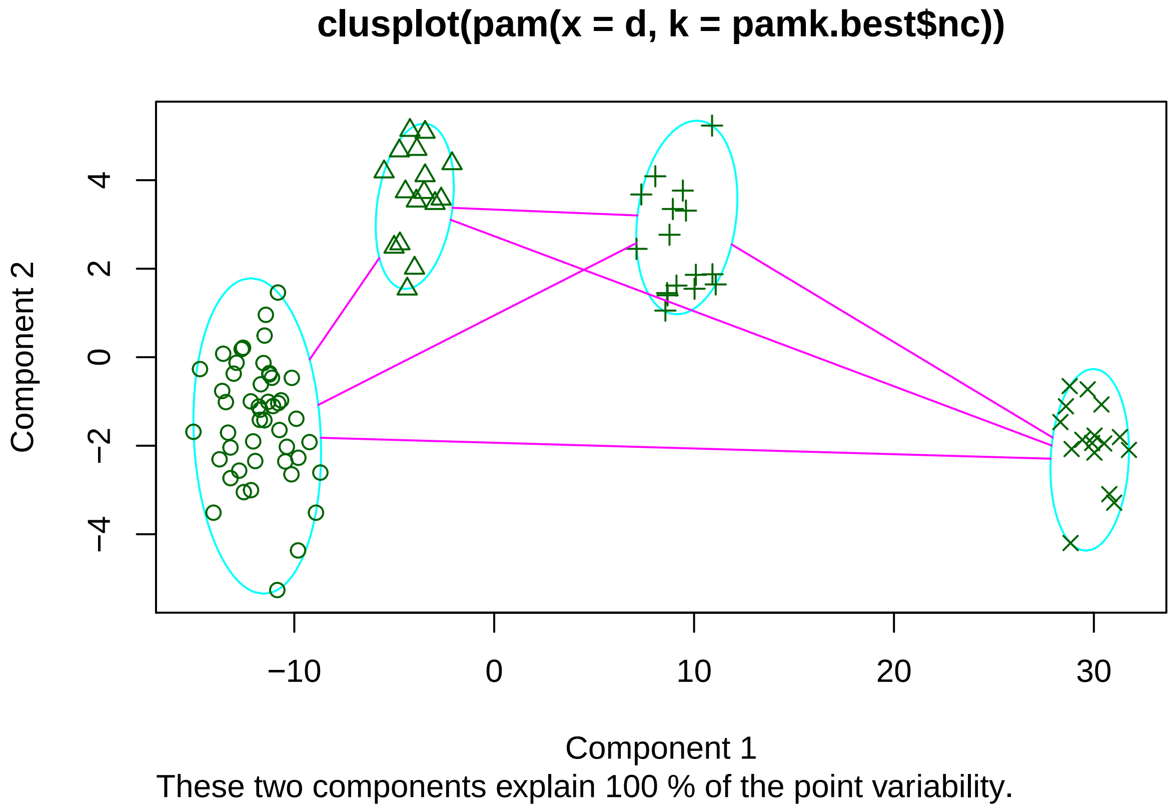
```r
# Sample data
n = 100
g = 6
set.seed(g)
d <- data.frame(x = unlist(lapply(1:g, function(i) rnorm(n/g, runif(1)*i^2))),
                y = unlist(lapply(1:g, function(i) rnorm(n/g, runif(1)*i^2))))
plot(d)
```

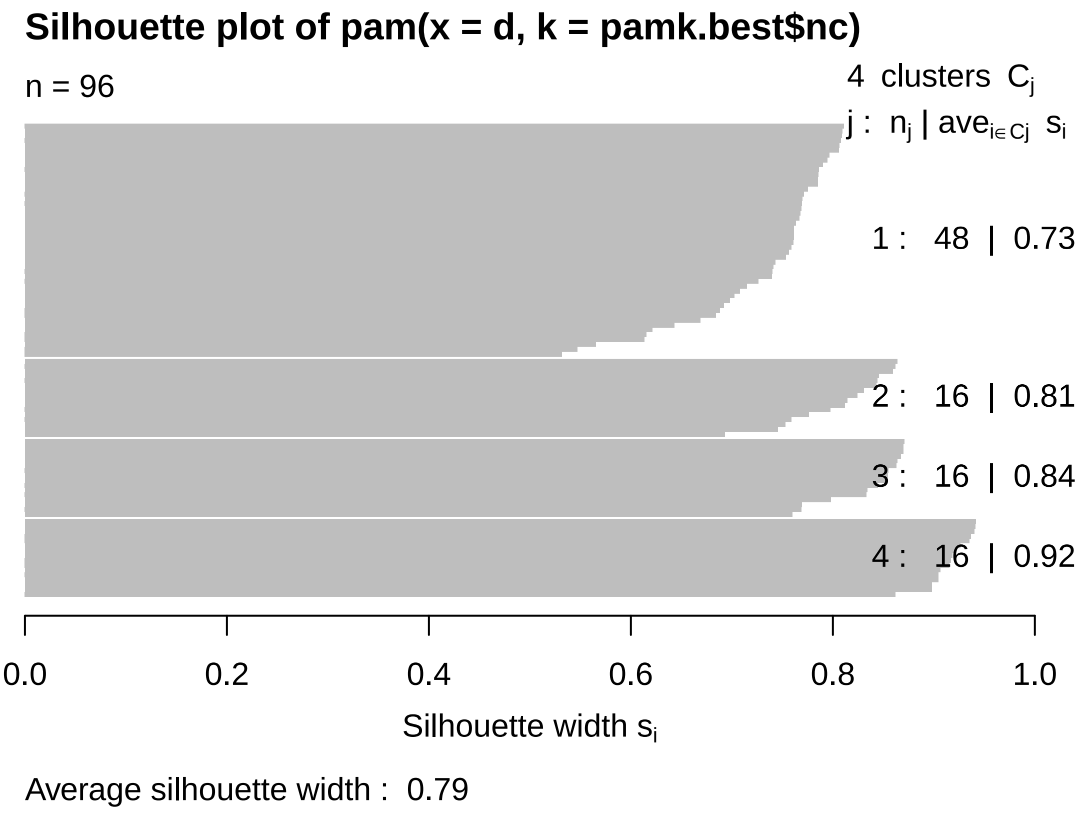


```
lots_of_clusters(d,4)
```



```
## number of clusters estimated by optimum average silhouette width: 4
```





```
## silhouette-optimal number of clusters: 4
```

```
# calinsky

# fit <- cascadeKM(scale(d, center = TRUE,  scale = TRUE), 1, 10, iter = 1000)
# plot(fit, sortg = TRUE, grpmts.plot = TRUE)
# calinski.best <- as.numeric(which.max(fit$results[2,]))
# cat("Calinski criterion optimal number of clusters:", calinski.best, "\n")
# BIC
# See http://www.jstatsoft.org/v18/i06/paper
# http://www.stat.washington.edu/research/reports/2006/tr504.pdf
#
# Run the function to see how many clusters
# it finds to be optimal, set it to search for
# at least 1 model and up 20.
# d_clust <- Mclust(as.matrix(d), G=1:20)
# m.best <- dim(d_clust$z)[2]
# cat("model-based optimal number of clusters:", m.best, "\n")
# # 4 clusters
# plot(d_clust)


# Affinity Propogation
# d.apclus <- apcluster(negDistMat(r=2), d)
# cat("affinity propogation optimal number of clusters:", length(d.apclus@clusters), "\n")
# # 4
# heatmap(d.apclus)
# plot(d.apclus, d)
# Gap Statistic


# clusGap(d, kmeans, 10, B = 100, verbose = interactive())

# NbClust
# nb <- NbClust(d, diss="NULL", distance = "euclidean",
#         min.nc=2, max.nc=15, method = "kmeans",
#         index = "alllong", alphaBeale = 0.1)
# hist(nb$Best.nc[1,], breaks = max(na.omit(nb$Best.nc[1,])))

# d_dist <- dist(as.matrix(d))   # find distance matrix
# plot(hclust(d_dist))           # apply hirarchical clustering and plot
#
# x <- as.matrix(d)
# d.bclus <- bclust(x, transformed.par = c(0, -50, log(16), 0, 0, 0))
# viplot(imp(d.bclus)$var); plot(d.bclus); ditplot(d.bclus)
# dptplot(d.bclus, scale = 20, horizbar.plot = TRUE,varimp = imp(d.bclus)$var, horizbar.distance = 0, d


# For high dimensional data
# data(Boston)
# boston.pv <- pvclust(Boston)
# plot(boston.pv)
```