Lab 18

quan le

2022-10-25

Tuning and Validation for Clustering

Consensus Clustering

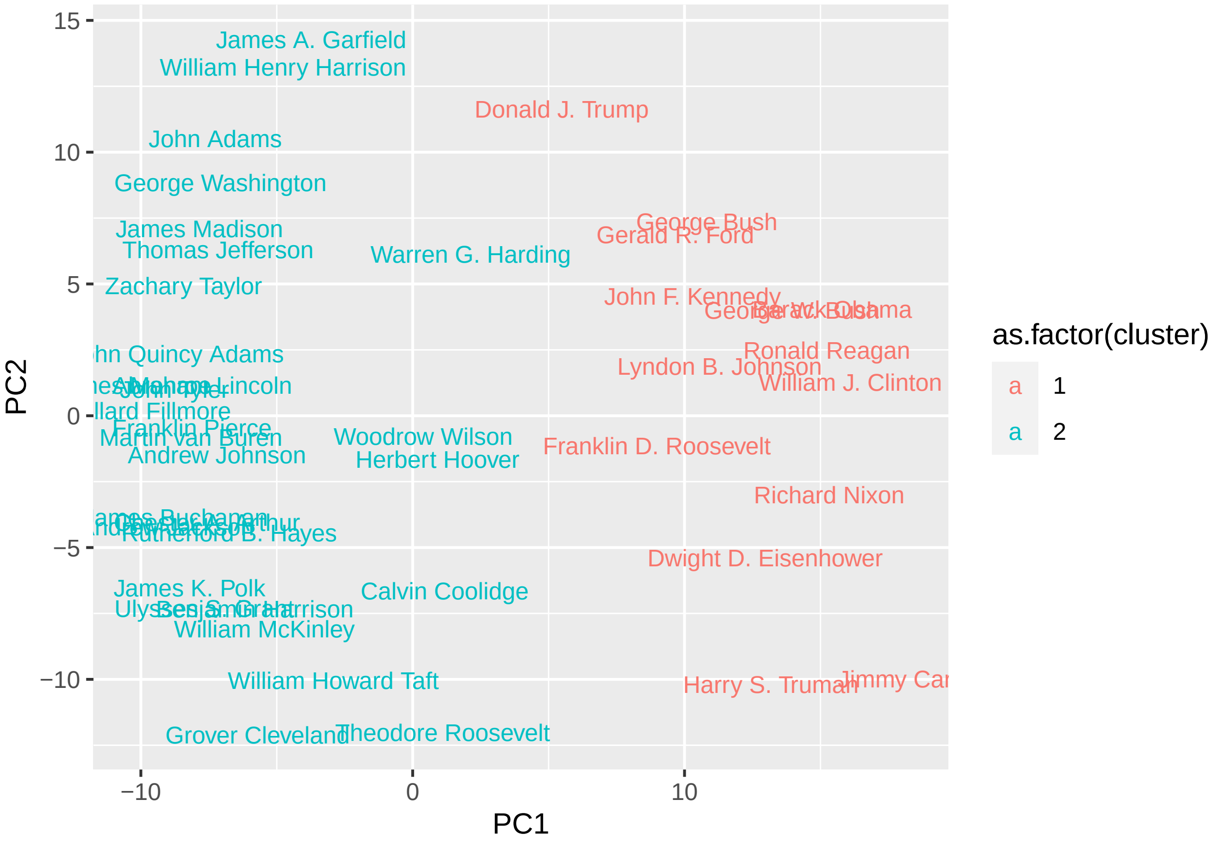
Data

```r
# data from https://github.com/DataSlingers/clustRviz/tree/master/data
load("data/presidential_speech.rda")
pdat = presidential_speech
```

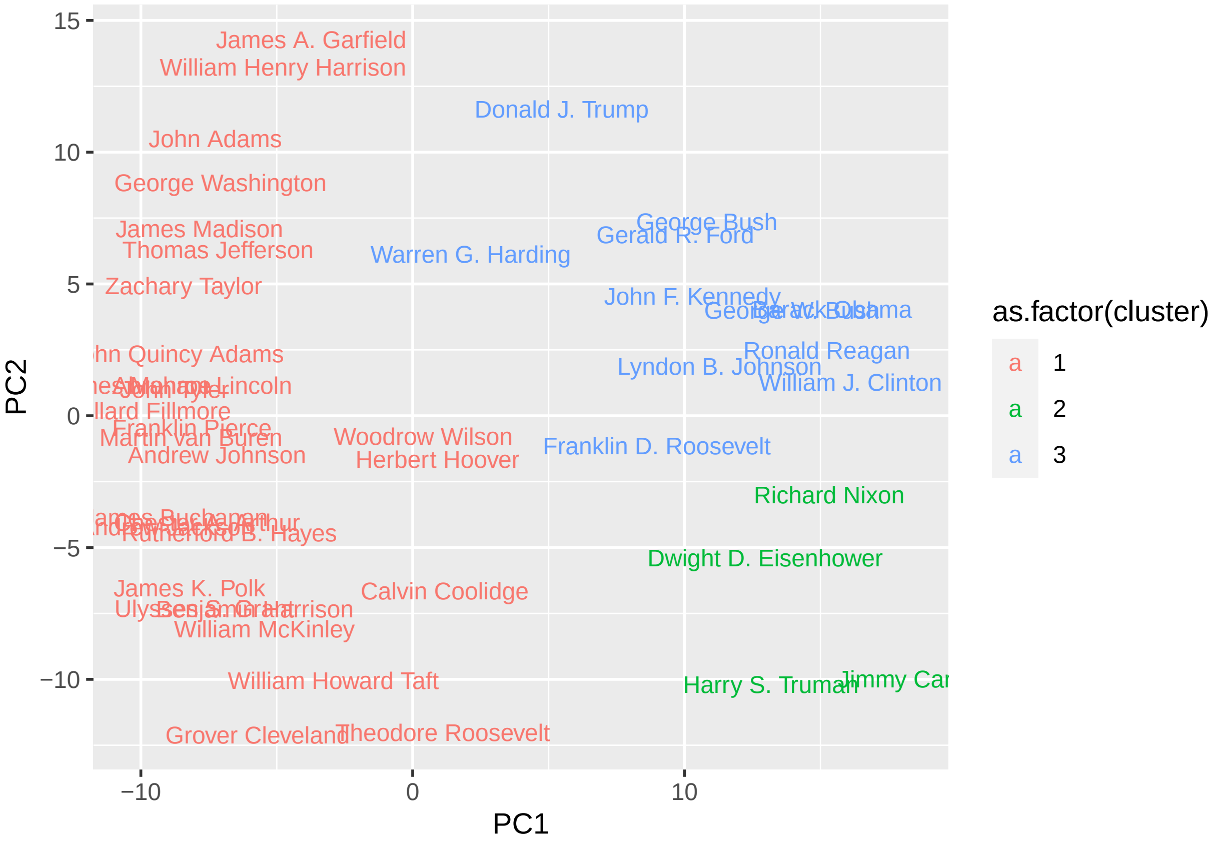
Get SVD for plotting purposes

```r
X = scale(pdat,center=TRUE,scale=FALSE)
sv = svd(X)
U = sv$u
V = sv$v
D = sv$d
Z = X%*%V
```

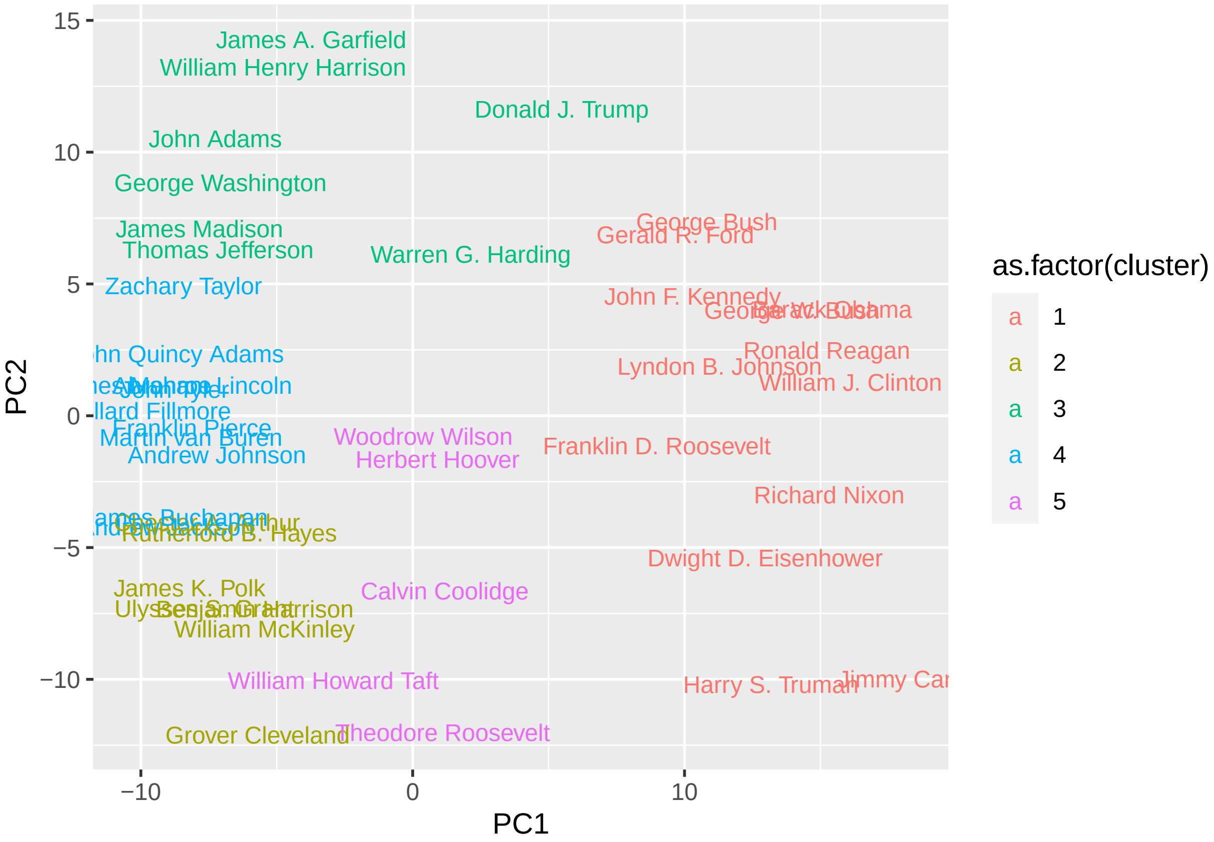
```r
K = 2
km = kmeans(X,centers=K)

clustered = data.frame(cbind(Z[,1],Z[,2],km$cluster),stringsAsFactors = FALSE)
colnames(clustered) = c("PC1","PC2","cluster")
clustered$PC1 = as.numeric(clustered$PC1)
clustered$PC2 = as.numeric(clustered$PC2)
# projected k-means centers
group.data = data.frame(km$centers%*%V[,1:2])
group.data$label = rownames(group.data)
colnames(group.data) = c("PC1","PC2","cluster")

ggplot(clustered,mapping=aes(x = PC1,y= PC2,color = as.factor(cluster))) +
  geom_text(mapping=aes(label = rownames(clustered)), size = 3)
```



```
K = 3
km = kmeans(X,centers=K)

clustered = data.frame(cbind(Z[,1],Z[,2],km$cluster),stringsAsFactors = FALSE)
colnames(clustered) = c("PC1","PC2","cluster")
clustered$PC1 = as.numeric(clustered$PC1)
clustered$PC2 = as.numeric(clustered$PC2)
# projected k-means centers
group.data = data.frame(km$centers%*%V[,1:2])
group.data$label = rownames(group.data)
colnames(group.data) = c("PC1","PC2","cluster")

ggplot(clustered,mapping=aes(x = PC1,y= PC2,color = as.factor(cluster))) +
  geom_text(mapping=aes(label = rownames(clustered)), size = 3)
```



```
K = 5
km = kmeans(X,centers=K)

clustered = data.frame(cbind(Z[,1],Z[,2],km$cluster),stringsAsFactors = FALSE)
colnames(clustered) = c("PC1","PC2","cluster")
clustered$PC1 = as.numeric(clustered$PC1)
clustered$PC2 = as.numeric(clustered$PC2)
# projected k-means centers
group.data = data.frame(km$centers%*%V[,1:2])
group.data$label = rownames(group.data)
colnames(group.data) = c("PC1","PC2","cluster")

ggplot(clustered,mapping=aes(x = PC1,y= PC2,color = as.factor(cluster))) +
  geom_text(mapping=aes(label = rownames(clustered)), size = 3)
```



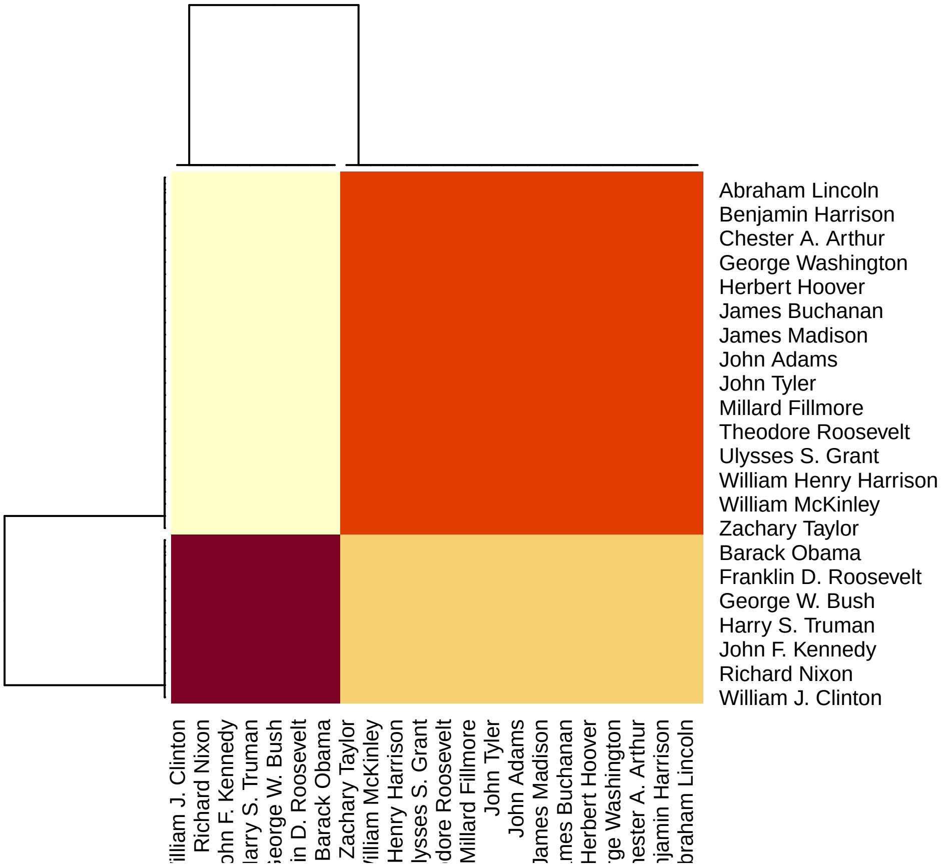
```
library(ConsensusClusterPlus)


results = ConsensusClusterPlus(t(pdat),maxK=6,reps=50,pItem=1,pFeature=1,
clusterAlg="km",plot="png")
```

```
## Note: The km (kmeans) option only supports a euclidean distance metric when supplying a data matrix.
```

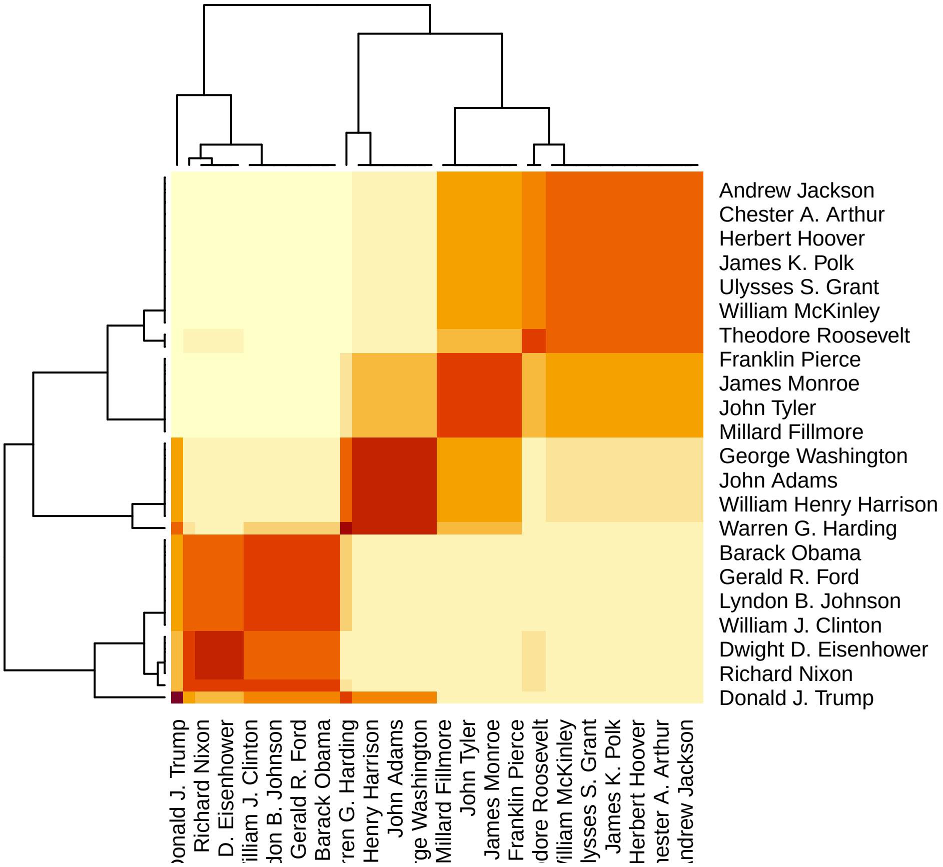
```
## end fraction
```

```
## clustered
## clustered
## clustered
## clustered
## clustered
```

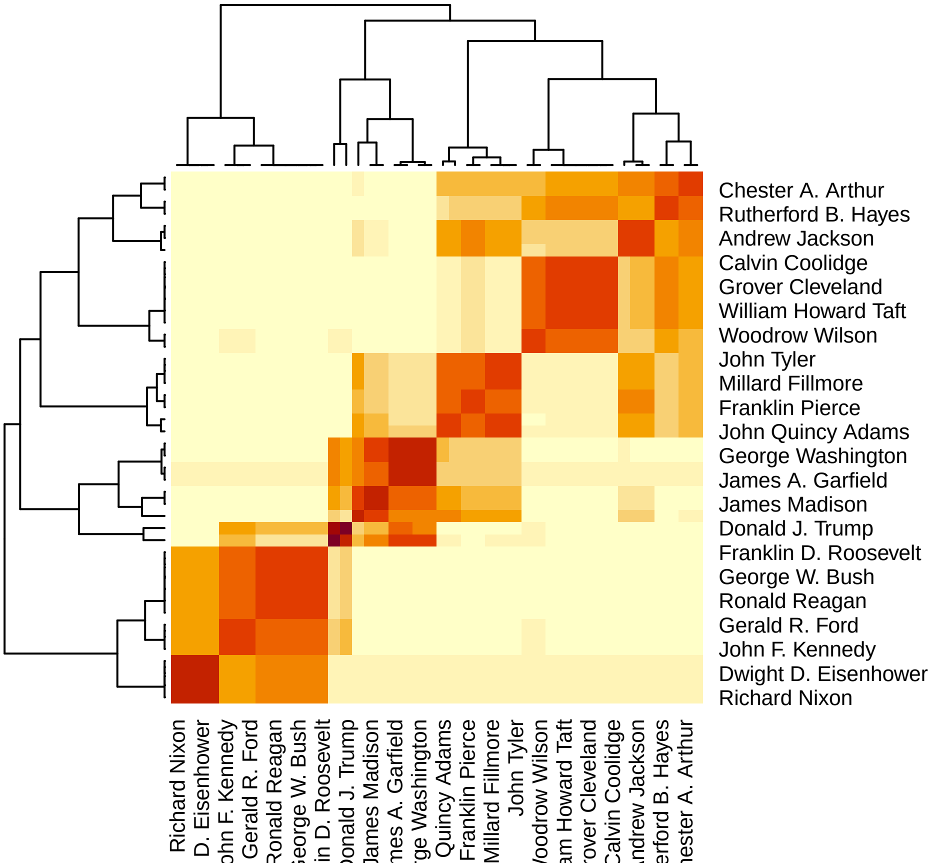
```
heatmap(results[[2]][["consensusMatrix"]], labRow=rownames(pdat), labCol=rownames(pdat))
```



```
heatmap(results[[3]][["consensusMatrix"]], labRow=rownames(pdat), labCol=rownames(pdat))
```



```
heatmap(results[[4]][["consensusMatrix"]], labRow=rownames(pdat), labCol=rownames(pdat))
```



Silhouette statistic

```
library(cluster)

K = 2
km = kmeans(X,centers=K)
sils = silhouette(km$cluster, dist(X))
rownames(sils) = names(km$cluster)
sils
```

```
##                       cluster neighbor  sil_width
## Abraham Lincoln             1        2 0.53174927
## Andrew Jackson              1        2 0.50964109
## Andrew Johnson              1        2 0.51392373
## Barack Obama                2        1 0.51439119
## Benjamin Harrison           1        2 0.45516436
## Calvin Coolidge             1        2 0.18258554
## Chester A. Arthur           1        2 0.51421949
## Donald J. Trump             2        1 0.15425676
## Dwight D. Eisenhower        2        1 0.41245098
## Franklin D. Roosevelt       2        1 0.31381263
## Franklin Pierce             1        2 0.51981793
## George Bush                 2        1 0.43700892
## George W. Bush              2        1 0.50733019
## George Washington           1        2 0.40089291
## Gerald R. Ford              2        1 0.37255697
## Grover Cleveland             1        2 0.38939393
## Harry S. Truman              2        1 0.30501530
## Herbert Hoover               1        2 0.16260834
## James A. Garfield            1        2 0.22188941
## James Buchanan               1        2 0.51835565
## James K. Polk                1        2 0.49283895
## James Madison                1        2 0.44258654
## James Monroe                 1        2 0.52213595
## Jimmy Carter                 2        1 0.39298983
## John Adams                   1        2 0.37373890
## John F. Kennedy              2        1 0.43034994
## John Quincy Adams            1        2 0.50857072
## John Tyler                   1        2 0.53586878
## Lyndon B. Johnson            2        1 0.48100211
## Martin van Buren             1        2 0.52286511
## Millard Fillmore             1        2 0.54324817
## Richard Nixon                2        1 0.49767396
## Ronald Reagan                2        1 0.55446670
## Rutherford B. Hayes          1        2 0.48554914
## Theodore Roosevelt           1        2 0.16120155
## Thomas Jefferson             1        2 0.43242752
## Ulysses S. Grant             1        2 0.48902575
## Warren G. Harding            1        2 0.03373109
## William Henry Harrison       1        2 0.26847110
## William Howard Taft          1        2 0.34119900
## William J. Clinton           2        1 0.53147571
## William McKinley             1        2 0.43710110
## Woodrow Wilson               1        2 0.20357003
## Zachary Taylor               1        2 0.47953780
## attr(,"Ordered")
## [1] FALSE
## attr(,"call")
## silhouette.default(x = km$cluster, dist = dist(X))
## attr(,"class")
## [1] "silhouette"
```

```
K = 4
km = kmeans(X,centers=K)
sils = silhouette(km$cluster, dist(X))
rownames(sils) = names(km$cluster)
sils
```

```
##                        cluster neighbor   sil_width
## Abraham Lincoln              3        1  0.38668963
## Andrew Jackson               3        1  0.15772942
## Andrew Johnson               3        1  0.23366109
## Barack Obama                 4        2  0.45294525
## Benjamin Harrison            1        3  0.26591903
## Calvin Coolidge              1        3  0.29473191
## Chester A. Arthur            1        3 -0.07229251
## Donald J. Trump              2        4  0.18154914
## Dwight D. Eisenhower         4        1  0.36739486
## Franklin D. Roosevelt        4        1  0.27411764
## Franklin Pierce              3        1  0.34069221
## George Bush                  4        2  0.25280060
## George W. Bush               4        2  0.44206983
## George Washington            2        3  0.04920296
## Gerald R. Ford               4        2  0.18922544
## Grover Cleveland             1        3  0.32258893
## Harry S. Truman              4        1  0.22446091
## Herbert Hoover               1        3  0.10030948
## James A. Garfield            2        3  0.40979232
## James Buchanan               3        1  0.17447403
## James K. Polk                1        3 -0.04519733
## James Madison                3        2  0.09606458
## James Monroe                 3        1  0.42001318
## Jimmy Carter                 4        1  0.34505866
## John Adams                   2        3  0.15098249
## John F. Kennedy              4        2  0.31356060
## John Quincy Adams            3        2  0.41890872
## John Tyler                   3        1  0.41969185
## Lyndon B. Johnson            4        2  0.44805312
## Martin van Buren             3        1  0.34502305
## Millard Fillmore             3        1  0.36146070
## Richard Nixon                4        1  0.47825423
## Ronald Reagan                4        2  0.52682203
## Rutherford B. Hayes          1        3  0.04165453
## Theodore Roosevelt           1        3  0.33112730
## Thomas Jefferson             3        2  0.12149774
## Ulysses S. Grant             1        3  0.13210000
## Warren G. Harding            2        3  0.18884258
## William Henry Harrison       2        3  0.34652239
## William Howard Taft          1        3  0.34966359
## William J. Clinton           4        2  0.51487724
## William McKinley             1        3  0.24149731
## Woodrow Wilson               1        3  0.05743819
## Zachary Taylor               3        2  0.24349503
## attr(,"Ordered")
## [1] FALSE
## attr(,"call")
## silhouette.default(x = km$cluster, dist = dist(X))
## attr(,"class")
## [1] "silhouette"
```

```r
set.seed(0)

for (K in 2:6) {
  silmeans = c()
  for (trial in 1:200) {
    km = kmeans(X,centers=K)
    sils = silhouette(km$cluster, dist(X))
    silmeans[trial] = mean(sils[,3])
  }
  print(paste("k =", K, "mean sil width", mean(silmeans)))
}
```

```
## [1] "k = 2 mean sil width 0.411333864362124"
## [1] "k = 3 mean sil width 0.322823529217111"
## [1] "k = 4 mean sil width 0.276626164556903"
## [1] "k = 5 mean sil width 0.246460695920333"
## [1] "k = 6 mean sil width 0.229231764459892"
```

Intro to Graphical Models

Load packages

```
library("igraph")
library("huge")
library("glasso")

library("glmnet")
```

```
load("data/sachs.Rdata")
p <- ncol(sachsdat)
n <- nrow(sachsdat)
sachscor <- cov2cor(sachscov)
```

Graphical lasso

Calculate lambda, based on formula in the slides (the third method)

```
alpha <- 0.01
num <- qt(p=alpha/(2*(p^2)),df=n-2, lower.tail=F)
lambda <- num / sqrt(n-2 + num)
```

Apply glasso

```
glasso.est <- glasso(s=sachscor,rho=lambda*4.2,approx=FALSE,
                     penalize.diagonal=FALSE)
A2 <- abs(glasso.est$wi) > 1E-16
diag(A2) <- 0
g2 <- graph.adjacency(A2, mode="undirected")
```

Neighborhood selection

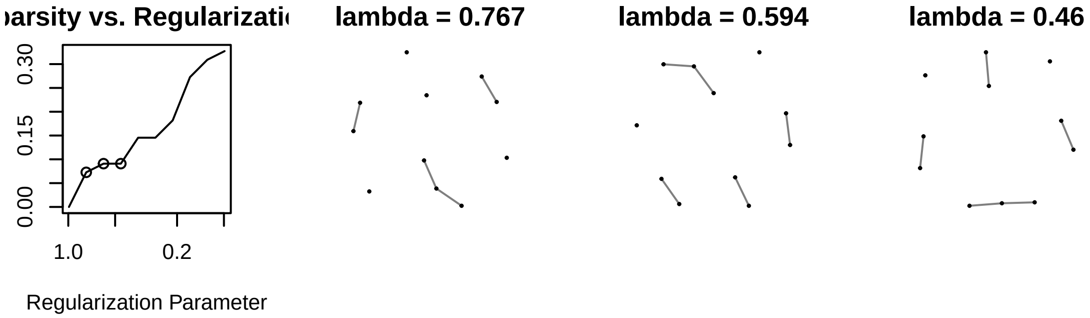
```
ns.est <- glasso(s=sachscor, rho=lambda, approx=TRUE, penalize.diagonal=FALSE)
A3 <- abs(ns.est$wi) > 1E-16; diag(A3) <- 0
g3 <- graph.adjacency(A3, mode="undirected")
```

Neighborhood selection estimate with huge (Stability selection for the value of λ)

```
X <- data.matrix(scale(sachsdat))
neth = huge(X,method="mb")
```

```
## Conducting Meinshausen & Buhlmann graph estimation (mb)....done
```

```
plot(neth)
```



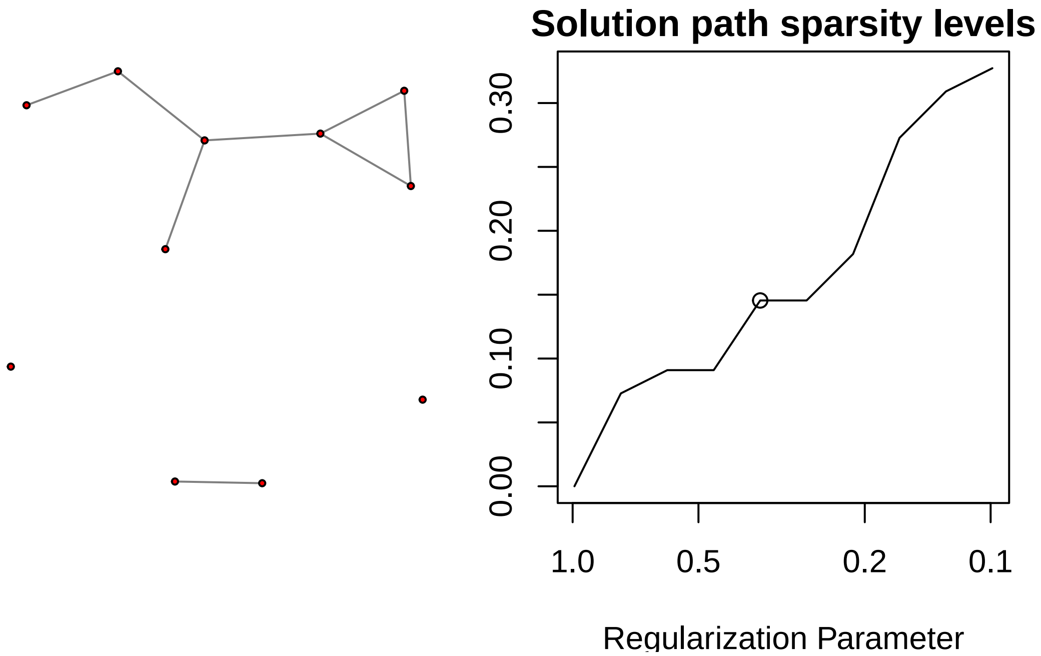
```
## stability selection with huge
net.s <- huge.select(neth, criterion="stars")
```

```
## Conducting Subsampling....in progress:5% Conducting Subsampling....in progress:10% Conducting Subsamp
```

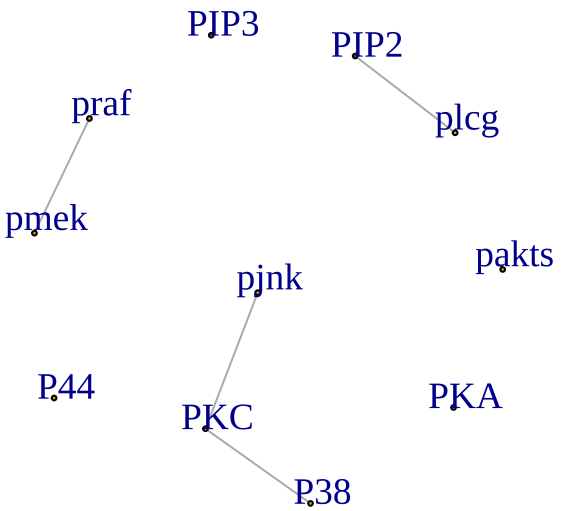
```
net.s
```

```
## Model: Meinshausen & Buhlmann Graph Estimation (mb)
## selection criterion: stars
## Graph dimension: 11
## sparsity level 0.1454545
```

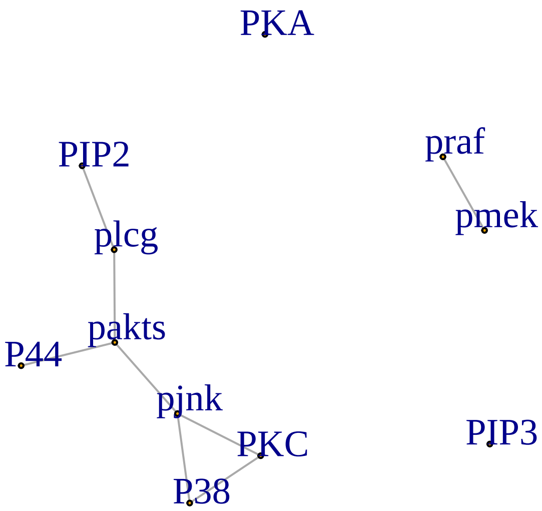
```
plot(net.s)
```



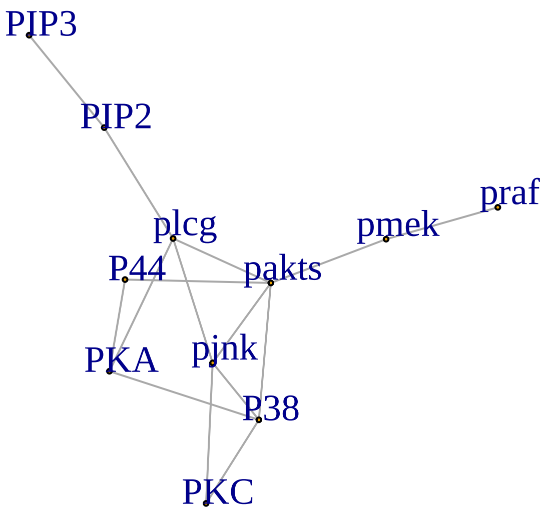
```
#larger lambda
mat <- neth$path[[2]]
neti <- as.undirected(graph_from_adjacency_matrix(mat))
plot(neti,vertex.label=colnames(X),vertex.size=2,vertex.label.cex=1.2,vertex.label.dist=1,layout=layout_
```



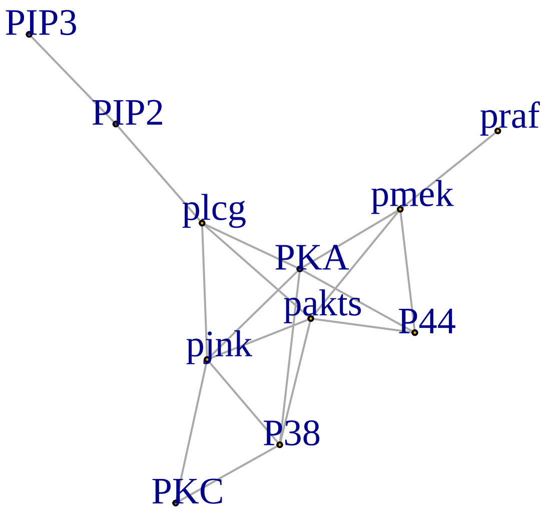
```
#smaller lambda
mat = neth$path[[5]]
neti = as.undirected(graph_from_adjacency_matrix(mat))
plot(neti,vertex.label=colnames(X),vertex.size=2,vertex.label.cex=1.2,vertex.label.dist=1,layout=layout_
```



```
# even smaller lambda
mat = neth$path[[8]]
neti = as.undirected(graph_from_adjacency_matrix(mat))
plot(neti,vertex.label=colnames(X),vertex.size=2,vertex.label.cex=1.2,vertex.label.dist=1,layout=layout_
```



```
# smallest lambda
mat = neth$path[[10]]
neti = as.undirected(graph_from_adjacency_matrix(mat))
plot(neti,vertex.label=colnames(X),vertex.size=2,vertex.label.cex=1.2,vertex.label.dist=1,layout=layout_
```


PIP3
PIP2
praf
pmek
plcg
PKA
pakts
P44
pjnk
P38
PKC