

June 2, 2022

The results below are generated from an R script.

```
## Install a package manager and packages
if (!require("pacman")) {
  install.packages("pacman")
}
pacman::p_load(Rfast, foreach, doParallel, mvnfast, rstudioapi)
current_path = rstudioapi::getActiveDocumentContext()$path
setwd(dirname(current_path))

pacman::p_load_gh("pkimes/sigclust2")
shc = get("shc", env = environment(shc))

source("sequential_function.R")

# k = 3 # number of clusters (3 or 10)
# uneven = FALSE #whether or not to have uneven weights
# distribution = 't' # t distribution or normal distribution?
# iterations = 50 # number of iterations

n = 500 # total number of samples
alpha = 0.05

if (distribution=='t'){
  distribution_name = 'True distribution components: t-distribution (df=3) mixture distribution'
  samplefunc <- function(n, mu, sigma, w){
    rmixt(n = n,mu = mus,sigma = sigmas,w = w,df = 3)
  }
}else{
  distribution_name = 'True distribution: Normal mixture distribution'
  samplefunc <- function(n, mu, sigma, w){
    rmixn(n=n, mu=mus, sigma=sigmas, w=w)
  }
}

# formulating d, delta (dimension and distance between clusters)
if (k == 10){
  a = c(2, 1, 2, 2, 2, 3, 2, 4, 2, 5, 2, 6, 2, 7, 2, 8, 2, 9) # dim2
  b = c(8, 1, 8, 2, 8, 3, 8, 4, 8, 5, 8, 6, 8, 7, 8, 8, 8, 9) # dim8
  d_delta = matrix(c(a, b) , ncol = 2, byrow = T)
} else if (k == 3){
  a = c(2, 1, 2, 2, 2, 3, 2, 4, 2, 5, 2, 6, 2, 7, 2, 8, 2, 9) # dim2
```

```

b = c(8, 1, 8, 2, 8, 3, 8, 4, 8, 5, 8, 6, 8, 7, 8, 8, 8, 9) # dim8
d_delta = matrix(c(a, b) , ncol = 2, byrow = T)
} else {
  stop("k != 3 or 10")
}

#weights
w = rep.int(1, k)
if (uneven){
  w[1] = 1 / 4
  w[2] = 1 / 2
}
w = w / sum(w)

K = floor(sqrt(n / 2)) #num clusters to test
K = min(K, 14L) # to ensure not estimating too many clusters

coresToUse = floor(detectCores() / 2) # cores to use

# function which creates data and performs one iteration
simulation <- function(iteration) {
  # simulate data
  set.seed(18 + iteration)
  data = samplefunc(n=n, mu=mus, sigma=sigmas, w=w)

  D1 = data[1:floor(n / 2), ]
  D2 = data[(floor(n / 2) + 1):n, ]

  # Estimate no.clusters
  Cluster_numbers = estimate.cluster.all(D1, D2, alpha, K)
  sigclust_splits = sum(shc(data, alpha = alpha)$nd_type == "sig")
  return(c(unlist(Cluster_numbers, use.names = F), sigclust_splits + 1L))
}

meanEstimate = matrix(nrow = nrow(d_delta), ncol = iterations)
medianEstimate = meanEstimate
meanEstimateI2 = meanEstimate
medianEstimateI2 = meanEstimate
AICEstimate = meanEstimate
BICEstimate = meanEstimate
sigclustEstimate = meanEstimate
RIFThierEstimate = meanEstimate

# For parallel computing
cl <- makeCluster(coresToUse) #not to overload computer
registerDoParallel(cl)

for (j in 1:nrow(d_delta)) {

```

```

d = d_delta[j, 1]
delta = d_delta[j, 2]

#sigmas = lapply(c(3,1,1), function(x) diag(x, nrow=d))
sigmas = lapply(rep.int(1, k), function(x)
  diag(x, nrow = d))

#mus = zeros(k, d)
#mus[1,1] = delta
#mus[2,2] = -delta
#mus[3,2] = delta
mus = outer(rep.int(1L, k), seq.int(d)) + delta * seq.int(0, k - 1L)

estimates <-
  foreach(
    i = 1:iterations,
    .combine = cbind,
    .inorder = F,
    .packages = c("mclust", "Rfast", "mvnfast", "MASS"),
    .verbose = F
  ) %dopar% {
    simulation(i)
  }

# format data into table
meanEstimate[j, ] = estimates[1, ]
medianEstimate[j, ] = estimates[2, ]
meanEstimate12[j, ] = estimates[3, ]
medianEstimate12[j, ] = estimates[4, ]
BICEstimate[j, ] = estimates[5, ]
AICEstimate[j, ] = estimates[6, ]
RIFTThierEstimate[j, ] = estimates[7, ]
sigclustEstimate[j, ] = estimates[8, ]
df = stack(data.frame(
  cbind(
    "Mean" = meanEstimate[j, ],
    "Mean12" = meanEstimate12[j, ],
    "Median" = medianEstimate[j, ],
    "Median12" = medianEstimate12[j, ],
    "AIC" = AICEstimate[j, ],
    "BIC" = BICEstimate[j, ],
    "RIFT.hc" = RIFTThierEstimate[j, ],
    "shc" = sigclustEstimate[j, ]
  )
))
print(paste0("(dimension, delta) = (", d, ",", delta, ")"))
colnames(df) = c("ESTIMATE", "METHOD")
tableEstimates = with(df, table(METHOD, ESTIMATE))
print(tableEstimates)
}

```

```

## [1] "(dimension, delta) = (2,1)"
## ESTIMATE
## METHOD      1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 19 20
## Mean       65 34  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## Meanl2     64 35  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## Median     33 59  8  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## Medianl2   33 59  8  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## AIC        15 66 18  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## BIC        52 48  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## RIFT.hc    18 24 39 17  2  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## shc         0  0  0  0  6  7  6  8 14 17  5 12 10  3  6  1  3  1  1
## [1] "(dimension, delta) = (2,2)"
## ESTIMATE
## METHOD      1  2  3  4  5  7  8  9 10 11 12 13 14 15 16 17
## Mean       52 41  6  1  0  0  0  0  0  0  0  0  0  0  0  0
## Meanl2     52 41  6  1  0  0  0  0  0  0  0  0  0  0  0  0
## Median     19 55 23  1  1  0  0  0  0  1  0  0  0  0  0  0
## Medianl2   18 56 24  1  0  0  0  0  0  1  0  0  0  0  0  0
## AIC         3 57 36  3  1  0  0  0  0  0  0  0  0  0  0  0
## BIC        32 63  5  0  0  0  0  0  0  0  0  0  0  0  0  0
## RIFT.hc     3 39 40 16  2  0  0  0  0  0  0  0  0  0  0  0
## shc         0  0  0  0  0  1  3 11 12 19 25  8 12  7  1  1
## [1] "(dimension, delta) = (2,3)"
## ESTIMATE
## METHOD      1  2  3  4  5  6  8  9 10 11 12 13 14 16
## Mean       28 20  7  0  2  0  0  8 11 15  6  3  0  0
## Meanl2     28 21  7  0  2  0  0  8 12 15  4  3  0  0
## Median      1  2  0  1  0  1  1 19 23 23 10  8 11  0
## Medianl2    1  2  0  1  0  1  1 18 25 23 10  7 11  0
## AIC         3 26 19  5  1  0  0  2 14 18  8  4  0  0
## BIC        31 61  8  0  0  0  0  0  0  0  0  0  0  0
## RIFT.hc     6 38 26 20  9  1  0  0  0  0  0  0  0  0
## shc         0  0  0  0  0  0  0  2 59 21 13  3  1  1
## [1] "(dimension, delta) = (2,4)"
## ESTIMATE
## METHOD      1  2  3  4  5  6  7  8  9 10 11 12 13 14
## Mean        1  0  1  2  1  0  2  2 23 16 19 18  9  6
## Meanl2       1  1  1  1  1  0  2  2 23 16 19 18  9  6
## Median       0  0  0  0  0  1  0  6 29 23 16 10 11  4
## Medianl2     0  0  0  0  0  1  0 11 25 23 14 11 11  4
## AIC          0  0  0  0  0  1  1  0  9 13 23 26 17 10
## BIC         21 29  4  0  0  0  0  0  3  9 14 14  5  1
## RIFT.hc      6 40 27 15 10  1  1  0  0  0  0  0  0  0
## shc          0  0  0  0  0  0  0  0  0 89 10  1  0  0
## [1] "(dimension, delta) = (2,5)"
## ESTIMATE
## METHOD      1  2  3  4  5  6  7  8  9 10 11 12 13 14
## Mean        1  0  0  0  0  1  1  8 16 19 23  9 13  9
## Meanl2       1  0  0  0  0  2  1  7 15 20 22 10 12 10
## Median       0  0  0  0  0  1  0 14 24 21 15 10  6  9
## Medianl2     0  0  0  0  0  1  0 15 21 23 14 10  8  8
## AIC          0  0  0  0  0  0  0  2 12 19 17 11 21 18
## BIC          0  4  1  0  0  0  0  0 15 18 20  9 20 13

```

```

## RIFT.hc 5 32 27 17 15 1 3 0 0 0 0 0 0 0 0
## shc 0 0 0 0 0 0 0 0 0 0 88 11 1 0 0
## [1] "(dimension, delta) = (2,6)"
## ESTIMATE
## METHOD 1 2 3 4 5 6 7 8 9 10 11 12 13 14
## Mean 0 0 0 1 0 0 2 7 14 12 12 19 15 18
## Meanl2 0 0 0 1 0 0 2 7 14 12 11 20 14 19
## Median 0 0 0 0 0 0 2 12 33 15 5 11 7 15
## Medianl2 0 0 0 0 0 0 2 13 29 19 6 10 7 14
## AIC 0 0 0 0 0 0 0 1 17 10 15 17 21 19
## BIC 0 0 0 0 0 0 1 1 20 10 14 16 20 18
## RIFT.hc 8 31 27 18 11 2 3 0 0 0 0 0 0 0
## shc 0 0 0 0 0 0 0 0 0 0 91 9 0 0 0
## [1] "(dimension, delta) = (2,7)"
## ESTIMATE
## METHOD 1 2 3 4 5 6 7 8 9 10 11 12 13 14
## Mean 0 0 0 0 0 0 0 7 17 7 8 17 20 24
## Meanl2 0 0 0 0 0 0 0 7 17 7 7 18 19 25
## Median 0 0 0 0 0 0 0 15 30 22 6 6 11 10
## Medianl2 0 0 0 0 0 0 0 15 29 22 6 8 9 11
## AIC 0 0 0 0 0 0 0 1 15 5 8 19 26 26
## BIC 0 0 0 0 0 0 0 2 21 7 7 17 20 26
## RIFT.hc 9 34 30 14 8 2 2 1 0 0 0 0 0 0
## shc 0 0 0 0 0 0 0 0 0 89 10 1 0 0
## [1] "(dimension, delta) = (2,8)"
## ESTIMATE
## METHOD 1 2 3 4 5 6 7 8 9 10 11 12 13 14
## Mean 0 0 0 0 0 0 0 2 16 7 13 24 18 20
## Meanl2 0 0 0 0 0 0 0 2 16 7 13 23 19 20
## Median 0 0 0 0 0 0 1 6 53 11 6 6 8 9
## Medianl2 0 0 0 0 0 0 1 10 47 12 7 6 8 9
## AIC 0 0 0 0 0 0 0 0 13 3 10 28 23 23
## BIC 0 0 0 0 0 0 0 0 15 3 11 26 22 23
## RIFT.hc 9 35 28 15 5 3 4 1 0 0 0 0 0 0
## shc 0 0 0 0 0 0 0 0 0 90 9 1 0 0
## [1] "(dimension, delta) = (2,9)"
## ESTIMATE
## METHOD 1 2 3 4 5 6 7 8 9 10 11 12 13 14
## Mean 0 0 0 0 0 0 0 1 15 12 17 23 18 14
## Meanl2 0 0 0 0 0 0 0 1 15 11 17 24 18 14
## Median 0 0 0 0 0 0 1 4 43 17 5 9 8 13
## Medianl2 0 0 0 0 0 0 1 8 38 16 7 9 8 13
## AIC 0 0 0 0 0 0 0 1 14 8 17 23 20 17
## BIC 0 0 0 0 0 0 0 1 12 10 17 24 19 17
## RIFT.hc 7 35 23 20 7 4 4 0 0 0 0 0 0 0
## shc 0 0 0 0 0 0 0 0 0 89 10 1 0 0
## [1] "(dimension, delta) = (8,1)"
## ESTIMATE
## METHOD 1 2 3 4 5 6 9 10 11 12 13 14 15 16 17 18 19 20 24
## Mean 81 18 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## Meanl2 81 18 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## Median 33 62 5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## Medianl2 33 62 5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## AIC 59 40 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

```

```

## BIC      100  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## RIFT.hc   21  56  8 11  3  1  0  0  0  0  0  0  0  0  0  0  0  0  0
## shc       0  0  0  0  0  0  0  4  6  8 18 10 10 18 10  7  3  2  3  1
## [1] "(dimension, delta) = (8,2)"
## ESTIMATE
## METHOD      1  2  3  4  5  6  8  9 10 11 12 13 14
## Mean      56 14  1  1  0  0  0 14  6  4  2  1  1
## Meanl2    56 14  1  1  0  0  0 14  6  4  2  1  1
## Median     0  1  0  0  0  0 11 49 22 11  4  1  1
## Medianl2   0  1  0  0  0  0 11 49 23 10  4  1  1
## AIC       31 33  1  0  0  0  0 20 10  3  2  0  0
## BIC       99  1  0  0  0  0  0  0  0  0  0  0  0
## RIFT.hc   21 63  3  6  5  2  0  0  0  0  0  0  0
## shc       0  0  0  0  0  0  0  0 73 21  4  2  0
## [1] "(dimension, delta) = (8,3)"
## ESTIMATE
## METHOD      1  2  3  5  8  9 10 11 12 13 14
## Mean      15  7  3  0  0 21 22 15  7  5  5
## Meanl2    15  7  3  0  0 21 22 15  7  5  5
## Median     0  0  0  0  2 42 29 16  7  2  2
## Medianl2   0  0  0  0  2 42 28 16  7  2  3
## AIC       5  5  0  0  0 31 25 17  7  5  5
## BIC      83  1  0  0  0  3  3  8  2  0  0
## RIFT.hc   20 64 15  1  0  0  0  0  0  0  0
## shc       0  0  0  0  0  0 86 12  2  0  0
## [1] "(dimension, delta) = (8,4)"
## ESTIMATE
## METHOD      1  2  3  4  5  7  8  9 10 11 12 13 14
## Mean      8  4  0  0  1  1  4 34 20  9  5 10  4
## Meanl2    8  4  0  0  1  1  4 34 20  9  5 10  4
## Median     0  0  0  0  0  0  1 52 25 16  3  2  1
## Medianl2   0  0  0  0  0  0  1 52 24 17  3  2  1
## AIC       4  5  0  0  0  0  0 38 23 12  6  8  4
## BIC      44  1  0  0  0  0  0 27  8  8  3  7  2
## RIFT.hc   17 66 14  1  2  0  0  0  0  0  0  0  0
## shc       0  0  0  0  0  0  0  0 86 12  2  0  0
## [1] "(dimension, delta) = (8,5)"
## ESTIMATE
## METHOD      1  2  3  4  5  7  8  9 10 11 12 13 14
## Mean      4  1  0  0  0  6 23 34  7  8  3  7  7
## Meanl2    4  1  0  0  0  6 23 34  7  8  3  7  7
## Median     0  0  0  0  0  0  6 69 13  5  1  2  4
## Medianl2   0  0  0  0  0  0  6 69 13  5  1  2  4
## AIC       1  1  0  0  0  0  0 55 14 12  3  7  7
## BIC      16  0  0  0  0  0  0 49 12  7  3  6  7
## RIFT.hc   20 58 18  1  3  0  0  0  0  0  0  0  0
## shc       0  0  0  0  0  0  0  0 88 11  1  0  0
## [1] "(dimension, delta) = (8,6)"
## ESTIMATE
## METHOD      1  2  3  4  5  7  8  9 10 11 12 13 14
## Mean      3  1  0  0  0 13 23 19  9  9  6  6 11
## Meanl2    3  1  0  0  0 13 23 19  9  9  6  6 11
## Median     0  0  0  0  0  0  3 69 16  9  2  1  0
## Medianl2   0  0  0  0  0  0  3 69 16  9  2  1  0

```

```

##      AIC      0  1  0  0  0  0  1 48 14 11  8  6 11
##      BIC      9  0  0  0  0  0  0 44 14 11  7  6  9
##      RIFT.hc 19 54 23  1  3  0  0  0  0  0  0  0  0
##      shc      0  0  0  0  0  0  0  0  0 88 11  1  0  0
## [1] "(dimension, delta) = (8,7)"
##      ESTIMATE
## METHOD      1  2  3  5  6  7  8  9 10 11 12 13 14
##      Mean      3  0  0  0  3 12 19  5 17 19  5 10  7
##      Mean12     3  0  0  0  3 12 19  5 17 19  5 10  7
##      Median     0  0  0  0  0  0  2 51 27 16  2  2  0
##      Median12   0  0  0  0  0  0  2 50 28 16  2  2  0
##      AIC        0  0  0  0  0  0  3 35 20 20  5 10  7
##      BIC        5  0  0  0  0  0  0 32 21 20  5 10  7
##      RIFT.hc    18 56 20  6  0  0  0  0  0  0  0  0  0
##      shc        0  0  0  0  0  0  0  0  0 88 11  1  0  0
## [1] "(dimension, delta) = (8,8)"
##      ESTIMATE
## METHOD      1  2  3  5  6  7  8  9 10 11 12 13 14
##      Mean      1  1  0  0  1  5 14  5 34 18  8  8  5
##      Mean12     1  1  0  0  1  5 14  5 34 18  8  8  5
##      Median     0  0  0  0  0  0  3 36 39 15  2  3  2
##      Median12   0  0  0  0  0  0  3 36 39 15  2  3  2
##      AIC        0  0  0  0  0  0  4 22 35 18  7  9  5
##      BIC        5  0  0  0  0  0  0 20 36 18  7  9  5
##      RIFT.hc    17 58 19  6  0  0  0  0  0  0  0  0  0
##      shc        0  0  0  0  0  0  0  0  0 89 10  1  0  0
## [1] "(dimension, delta) = (8,9)"
##      ESTIMATE
## METHOD      1  2  3  5  6  7  8  9 10 11 12 13 14
##      Mean      1  0  0  0  1  5  6  3 44 21 10  3  6
##      Mean12     1  0  0  0  1  5  6  3 44 21 10  3  6
##      Median     0  0  0  0  0  0  2 26 46 16  6  2  2
##      Median12   0  0  0  0  0  0  2 26 46 16  6  2  2
##      AIC        0  0  0  0  0  0  2 11 44 24 10  3  6
##      BIC        3  0  0  0  0  0  0 12 44 22 10  3  6
##      RIFT.hc    20 58 16  5  1  0  0  0  0  0  0  0  0
##      shc        0  0  0  0  0  0  0  0  0 88 11  1  0  0

#stop cluster (parallel computing)
stopCluster(cl)

print(distribution_name )

## [1] "True distribution: Normal mixture distribution"

print(paste(k, 'true clusters:'))

## [1] "10 true clusters:"

print('Cluster weights:')

## [1] "Cluster weights:"

print(w)

## [1] 0.02857143 0.05714286 0.11428571 0.11428571 0.11428571 0.11428571 0.11428571
## [8] 0.11428571 0.11428571 0.11428571

```

The R session information (including the OS info, R version and all packages used):

```
sessionInfo()

## R version 4.1.2 (2021-11-01)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Monterey 12.0.1
##
## Matrix products: default
## LAPACK: /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_GB.UTF-8/en_GB.UTF-8/en_GB.UTF-8/C/en_GB.UTF-8/en_GB.UTF-8
##
## attached base packages:
## [1] grid      parallel  stats      graphics  grDevices  utils      datasets  methods
## [9] base
##
## other attached packages:
## [1] arm_1.12-2      lme4_1.1-27.1    Matrix_1.3-4     knitr_1.37
## [5] sigclust_1.1.0  mixtools_1.2.0   gridExtra_2.3     ggplot2_3.3.5
## [9] MASS_7.3-54     pracma_2.3.6     mclust_5.4.9     sigclust2_1.2.4
## [13] rstudioapi_0.13 mvnfast_0.2.7    doParallel_1.0.16 iterators_1.0.13
## [17] foreach_1.5.1   Rfast_2.0.6      RcppZigurat_0.1.6 Rcpp_1.0.8
## [21] pacman_0.5.1
##
## loaded via a namespace (and not attached):
## [1] minqa_1.2.4      colorspace_2.0-2    ellipsis_0.3.2
## [4] dynamicTreeCut_1.63-1 htmlTable_2.4.0     XVector_0.34.0
## [7] base64enc_0.1-3  gg dendro_0.1.23    bit64_4.0.5
## [10] AnnotationDbi_1.56.2 fansi_0.5.0         codetools_0.2-18
## [13] splines_4.1.2    cachem_1.0.6        impute_1.68.0
## [16] Formula_1.2-4    nloptr_1.2.2.3      broom_0.7.12
## [19] WGCNA_1.70-3     cluster_2.1.2       kernlab_0.9-29
## [22] GO.db_3.14.0     png_0.1-7           compiler_4.1.2
## [25] httr_1.4.2       backports_1.4.1     fastmap_1.1.0
## [28] htmltools_0.5.2  tools_4.1.2         coda_0.19-4
## [31] gtable_0.3.0     glue_1.6.1          GenomeInfoDbData_1.2.7
## [34] dplyr_1.0.7      ggthemes_4.2.4      Biobase_2.54.0
## [37] vctrs_0.4.1      Biostrings_2.62.0   preprocessCore_1.56.0
## [40] nlme_3.1-153     xfun_0.30           fastcluster_1.2.3
## [43] stringr_1.4.0    lifecycle_1.0.1     zlibbioc_1.40.0
## [46] scales_1.1.1     RColorBrewer_1.1-2  yaml_2.3.4
## [49] memoise_2.0.1    rpart_4.1-15        segmented_1.3-4
## [52] latticeExtra_0.6-29 stringi_1.7.6        RSQLite_2.2.10
## [55] highr_0.9        S4Vectors_0.32.3    blme_1.0-5
## [58] checkmate_2.0.0  BiocGenerics_0.40.0 boot_1.3-28
## [61] GenomeInfoDb_1.30.1 rlang_1.0.2         pkgconfig_2.0.3
## [64] matrixStats_0.61.0 bitops_1.0-7        evaluate_0.15
## [67] lattice_0.20-45  purrr_0.3.4         htmlwidgets_1.5.4
## [70] bit_4.0.4        tidyselect_1.1.1    magrittr_2.0.2
## [73] R6_2.5.1         IRanges_2.28.0      generics_0.1.1
## [76] Hmisc_4.6-0      DBI_1.1.2           pillar_1.6.4
## [79] foreign_0.8-81   withr_2.4.3         survival_3.2-13
## [82] KEGGREST_1.34.0  abind_1.4-5         RCurl_1.98-1.6
```



```
## [85] nnet_7.3-16          tibble_3.1.6          crayon_1.4.2
## [88] utf8_1.2.2           rmarkdown_2.13        jpeg_0.1-9
## [91] data.table_1.14.2    blob_1.2.2            forcats_0.5.1
## [94] digest_0.6.29        tidyr_1.1.4           stats4_4.1.2
## [97] munsell_0.5.0

Sys.time()

## [1] "2022-06-02 15:03:46 BST"
```