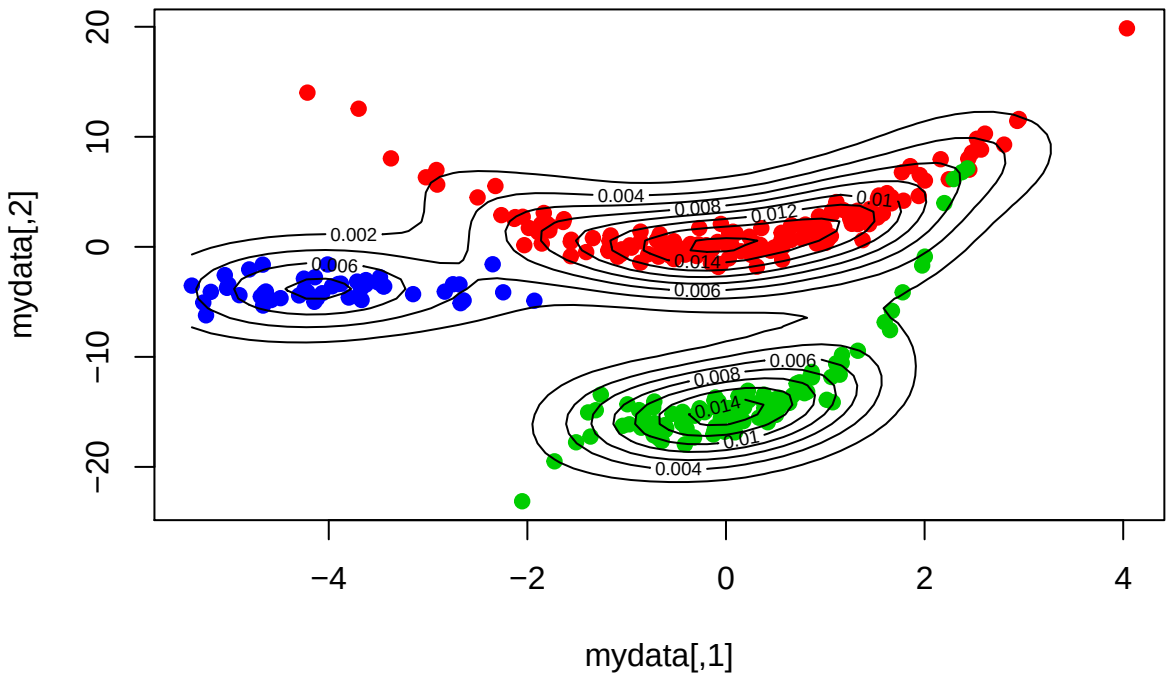
Cluster Comparison Simulation Study Supplementary

Jeremy Ash

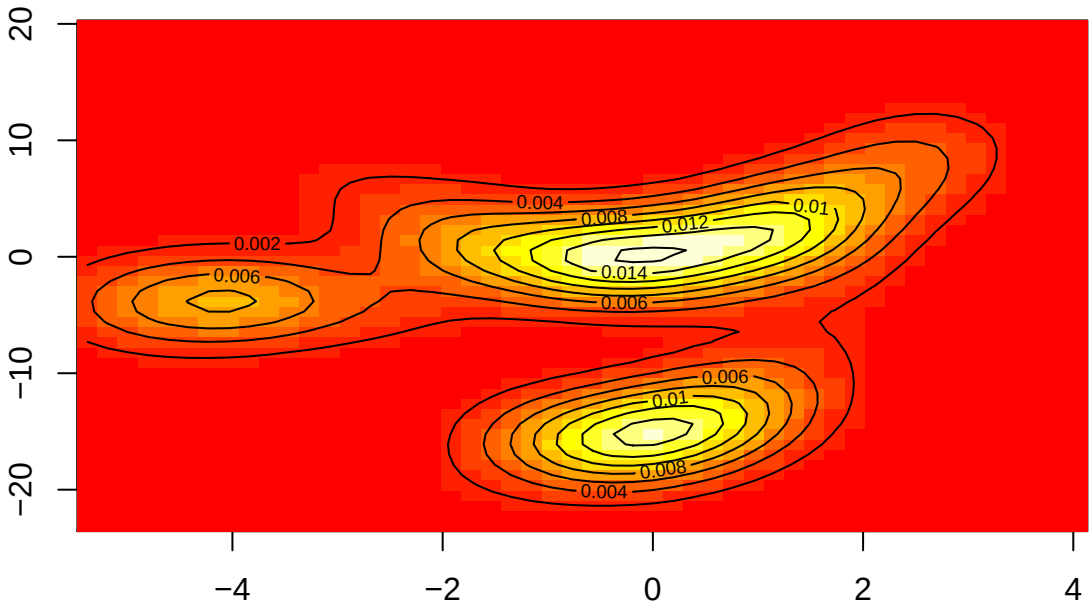
May 21, 2018

Simulation 1

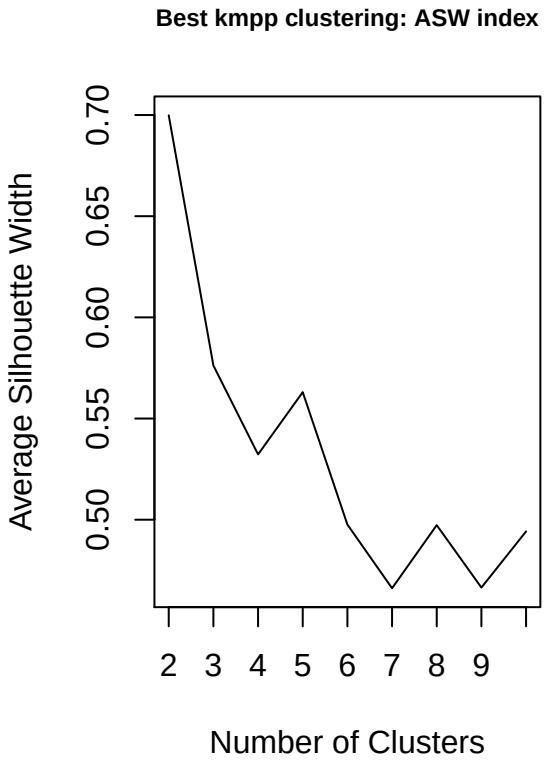
```
set.seed(123)

N <- 300

components <- sample(1:3,prob=c(15/30, 10/30, 5/30),size=N,replace=TRUE)

mydata <- matrix(ncol = 2, nrow = N)
for(i in 1:N) {
  if(components[i] == 2) {
    x <- rnorm(1)
    e <- rnorm(1)
    y = -15 + x + x^2 + x^3 + e
    mydata[i, ] <- c(x, y)
  } else if(components[i] == 1) {
    x <- rnorm(1, sd = 1.5)
    e <- rnorm(1)
    y = x + x^2 + e
    mydata[i, ] <- c(x, y)
  } else if(components[i] == 3){
    mydata[i, ] <- mvrnorm(1, c(-4,-4), matrix(c(1,0,0,1), ncol=2))
  }
}

truth <- components

plot(mydata, pch=19, col = truth + 1)
bivn.kde <- kde2d(mydata[,1], mydata[,2], n = 50)
contour(bivn.kde, add = TRUE)
```

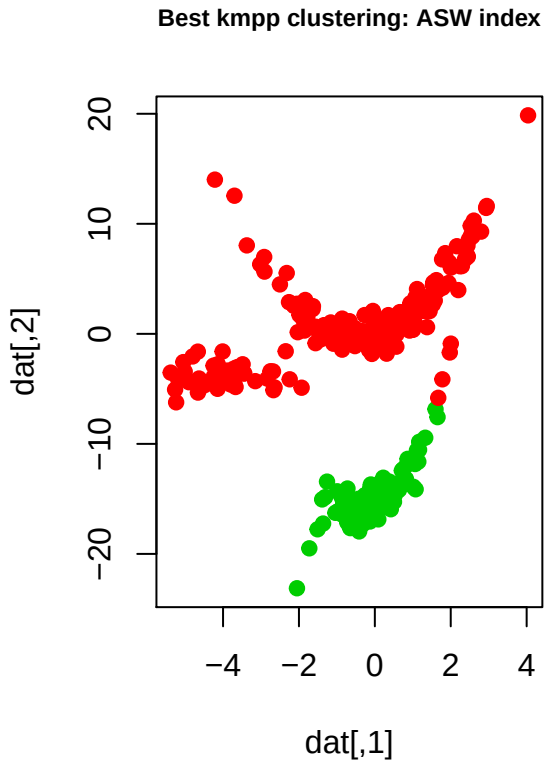


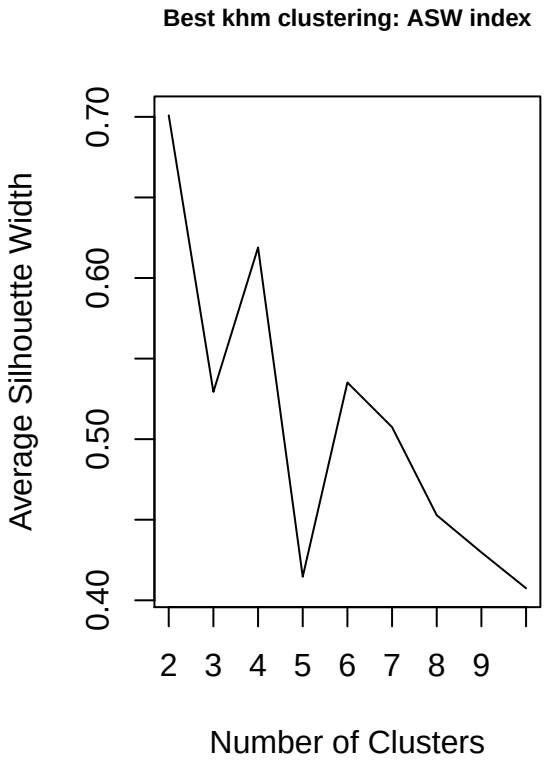

```
image(bivn.kde)       # from base graphics package
contour(bivn.kde, add = TRUE)
```

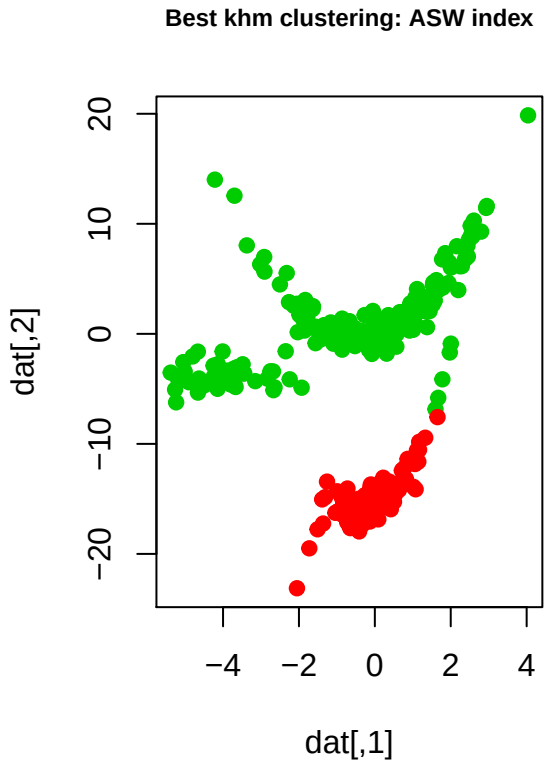


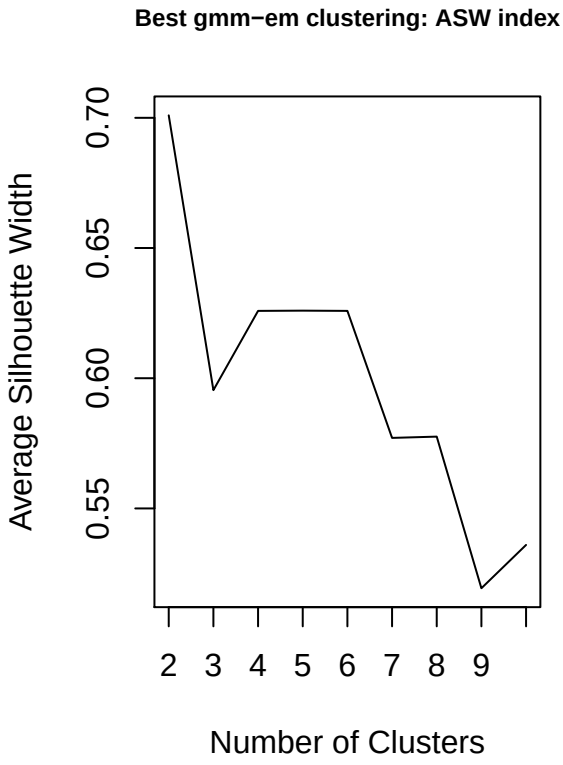

```
df <- simulate(mydata, truth, run.seed = 132, imax = 100, ks = seq(2, 10, 1))
```

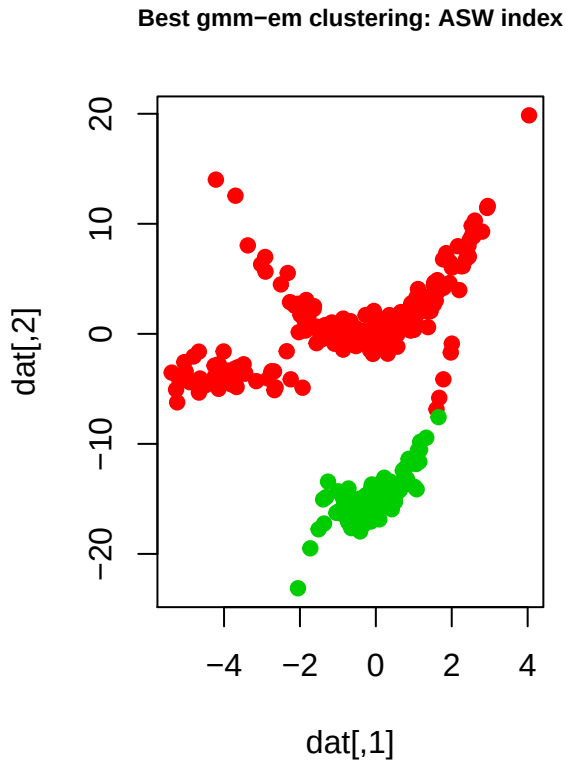

Best kmpp clustering: ASW index
Average Silhouette Width
0.50
0.55
0.60
0.65
0.70
2
3
4
5
6
7
8
9
Number of Clusters

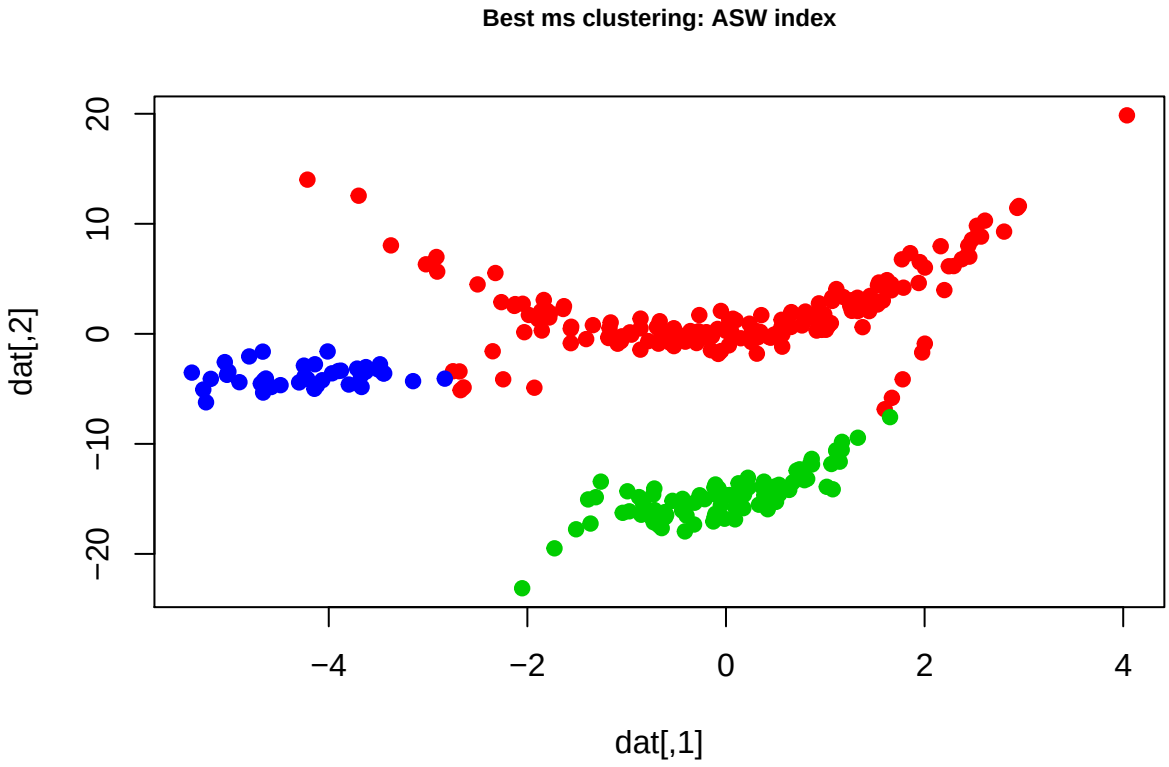

Best kmpp clustering: ASW index
dat[,2]
−20
−10
0
10
20
−4
−2
0
2
4
dat[,1]

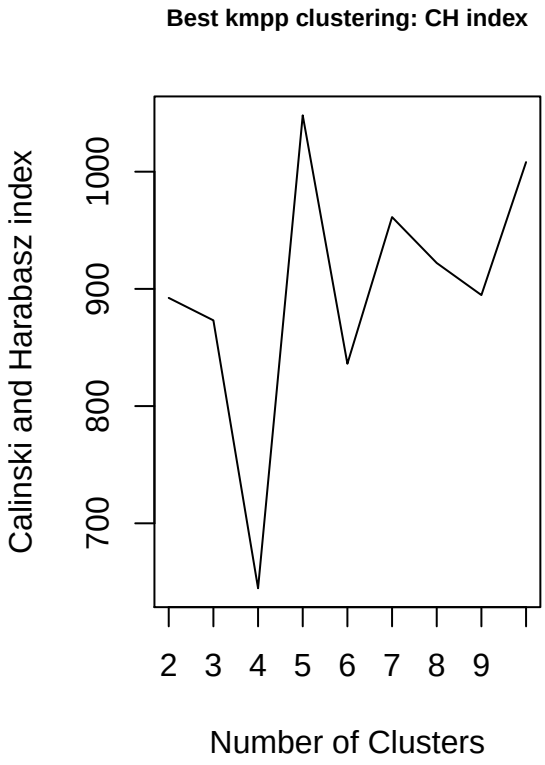

Best khm clustering: ASW index
Average Silhouette Width
0.40
0.50
0.60
0.70
2 3 4 5 6 7 8 9
Number of Clusters

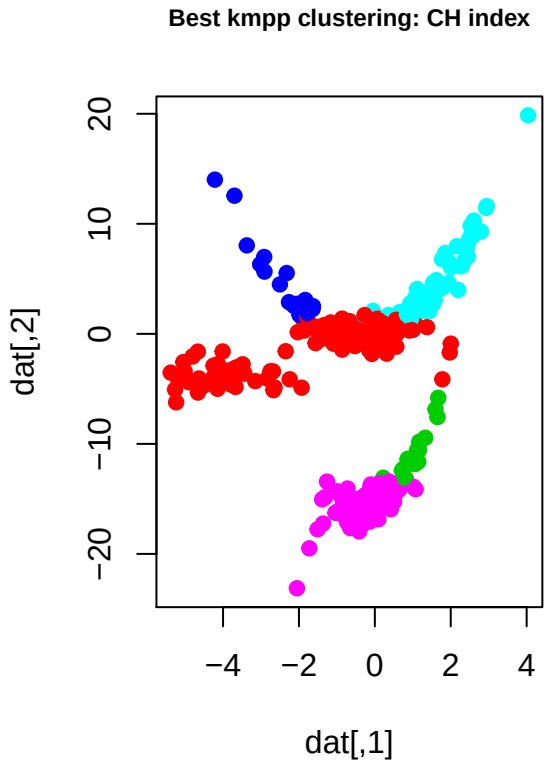

Best khm clustering: ASW index
dat[,2]
−20
−10
0
10
20
−4 −2 0 2 4
dat[,1]

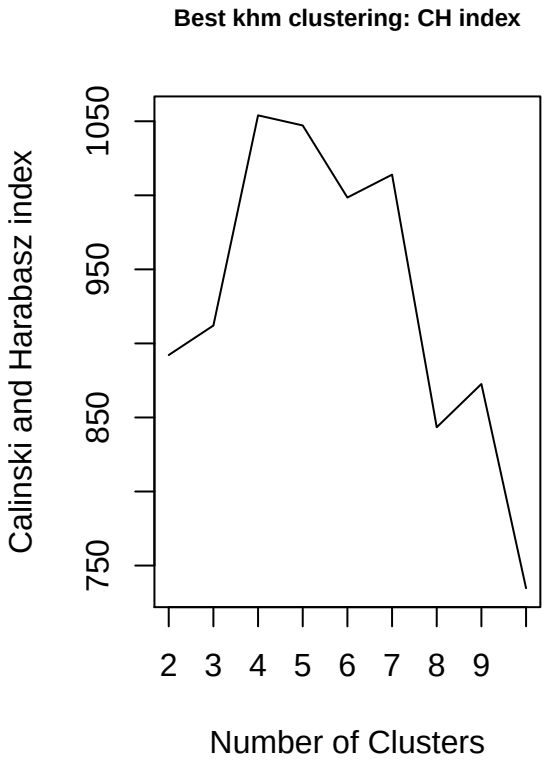

Best gmm−em clustering: ASW index
Average Silhouette Width
0.55
0.60
0.65
0.70
2 3 4 5 6 7 8 9
Number of Clusters

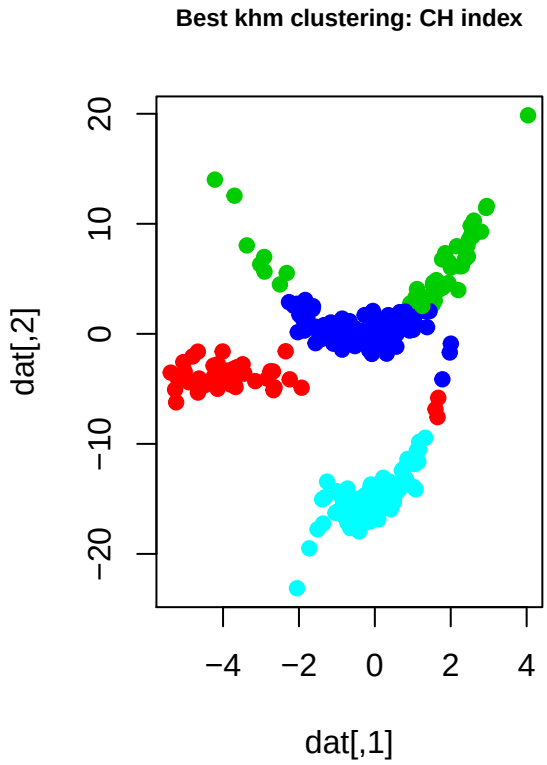

Best gmm−em clustering: ASW index
dat[,2]
−20
−10
0
10
20
−4 −2 0 2 4
dat[,1]

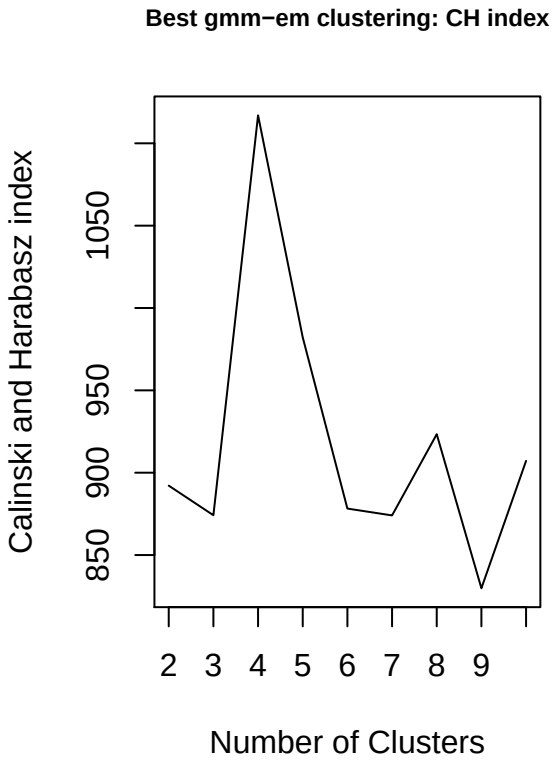

Best ms clustering: ASW index
dat[,2]
20
10
0
−10
−20
−4
−2
0
2
4
dat[,1]

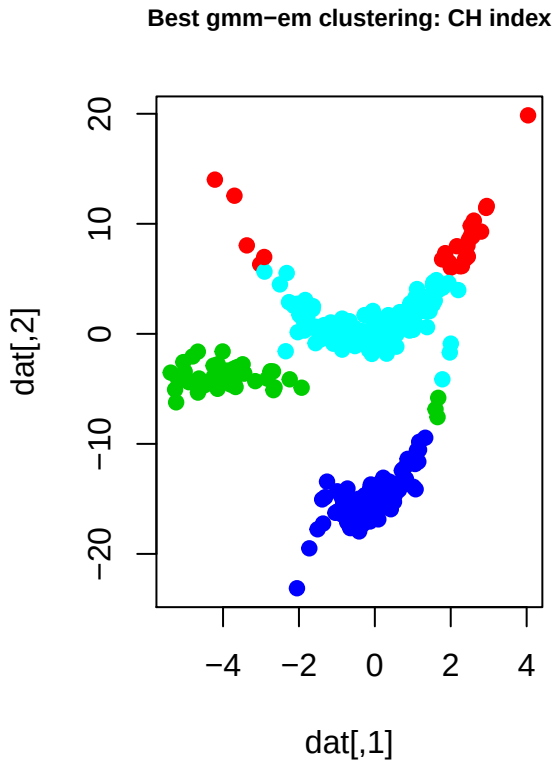

Best kmpp clustering: CH index
Calinski and Harabasz index
700
800
900
1000
2
3
4
5
6
7
8
9
Number of Clusters


Best kmpp clustering: CH index
dat[,2]
−20
−10
0
10
20
−4
−2
0
2
4
dat[,1]


Best khm clustering: CH index
Calinski and Harabasz index
750
850
950
1050
2 3 4 5 6 7 8 9
Number of Clusters


Best khm clustering: CH index
dat[,2]
−20
−10
0
10
20
−4 −2 0 2 4
dat[,1]


Best gmm−em clustering: CH index
Calinski and Harabasz index
850
900
950
1050
2
3
4
5
6
7
8
9
Number of Clusters


Best gmm−em clustering: CH index
dat[,2]
−20
−10
0
10
20
−4
−2
0
2
4
dat[,1]

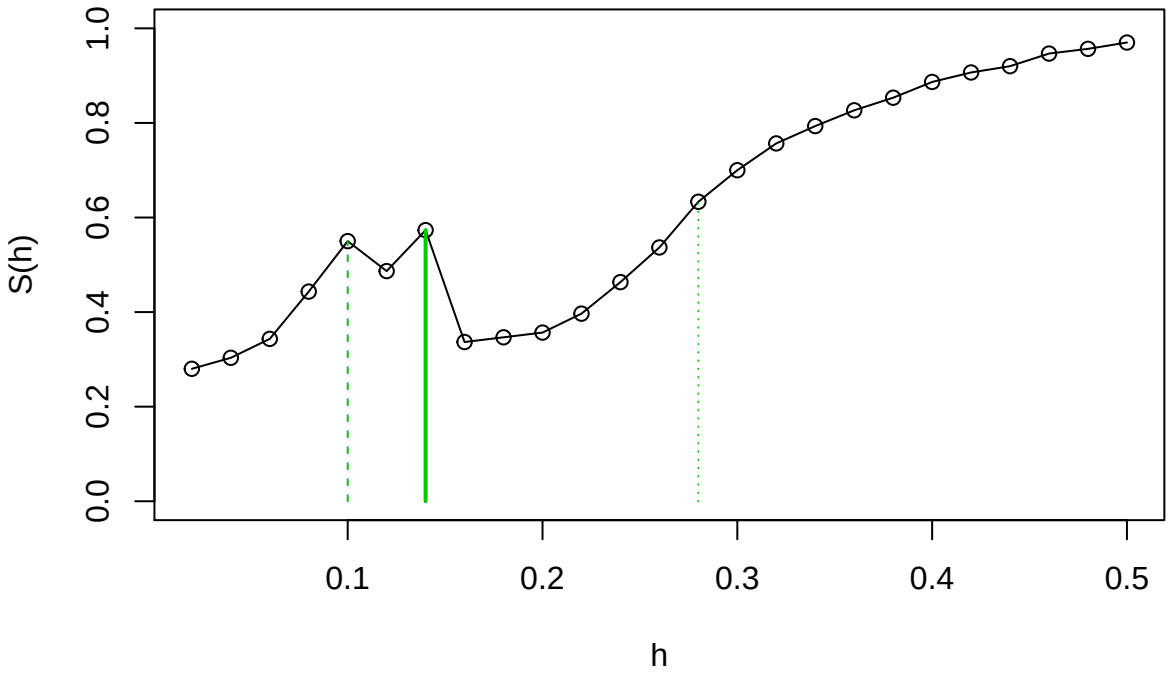
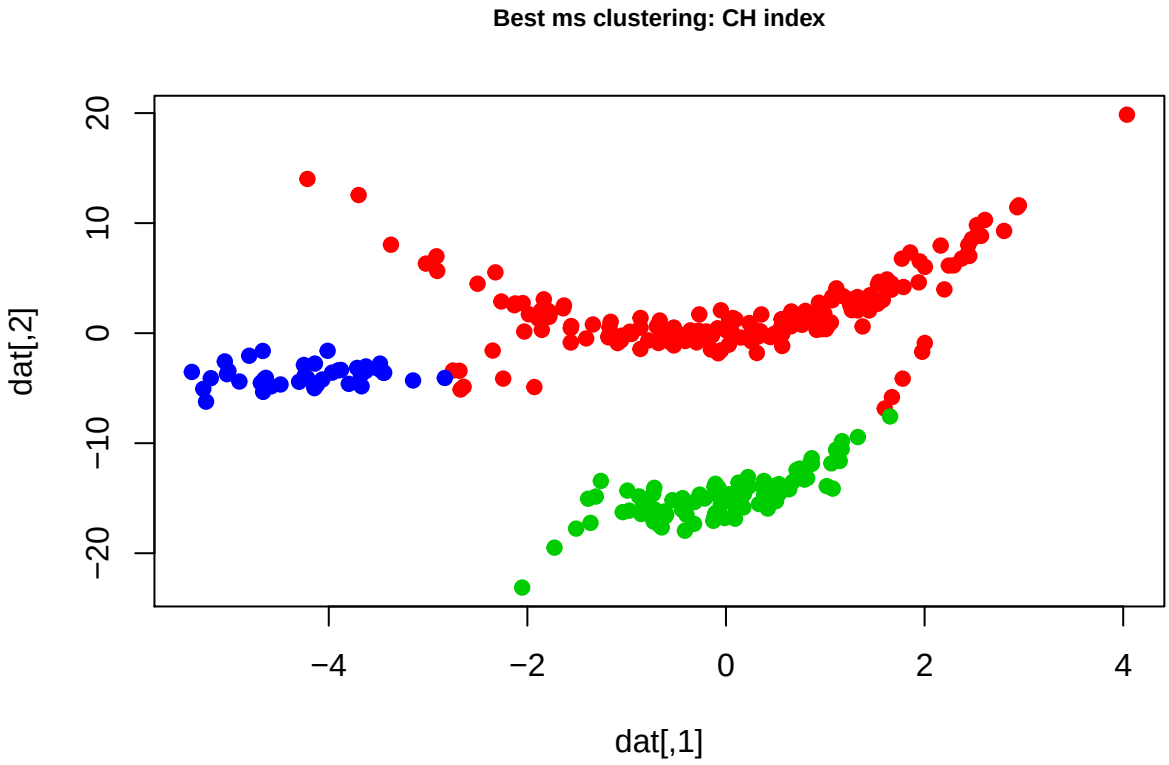

S(h)
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.2
0.3
0.4
0.5
h

Table 1: External validation measures for simulation data set 1. Number of clusters, k = 2, ..., 10, was selected by average silhouette width or Calinski and Harabasz index. The best performing method in bold.

	Average Silhouette Width							*Calinski and Harabasz index*						
	RI	HA	MA	FM	JI	VI	NMI	RI	HA	MA	FM	JI	VI	NMI
k-means++	0.78	0.58	0.58	0.79	0.63	0.83	0.64	0.70	0.34	0.35	0.57	0.39	1.63	0.52
k-harmonic means	0.78	0.57	0.57	0.79	0.63	0.86	0.63	0.85	0.67	0.67	0.79	0.63	0.86	0.74
GMM-EM	0.78	0.57	0.57	0.79	0.63	0.86	0.63	0.89	0.77	0.77	0.86	0.74	0.75	0.77
Mean Shift	**0.92**	**0.83**	**0.83**	**0.90**	**0.82**	**0.55**	**0.81**	**0.92**	**0.83**	**0.83**	**0.90**	**0.82**	**0.55**	**0.81**



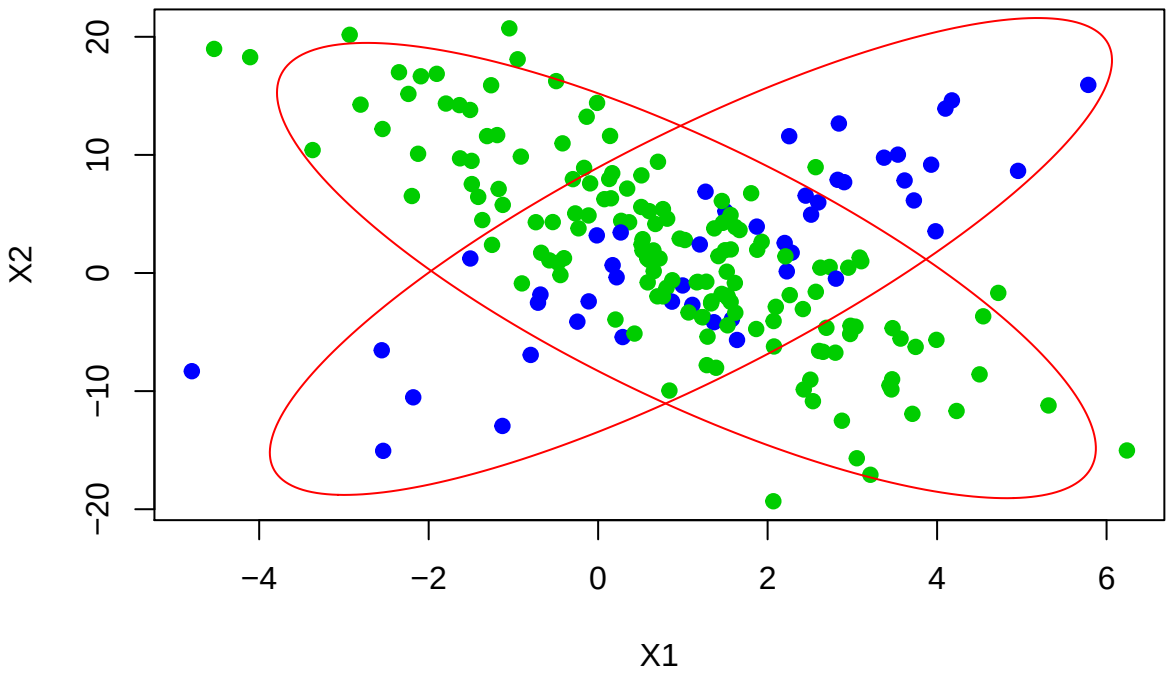

```
dat1 <- mydata
truth1 <- truth
```

```
df <- round(df, 2)
rownames(df) <- c("k-means++", "k-harmonic means", "GMM-EM", "Mean Shift")

kable(df, format = "latex", align = "c", booktabs = T, caption = "External validation measures for simul
row_spec(4, bold = T)
```

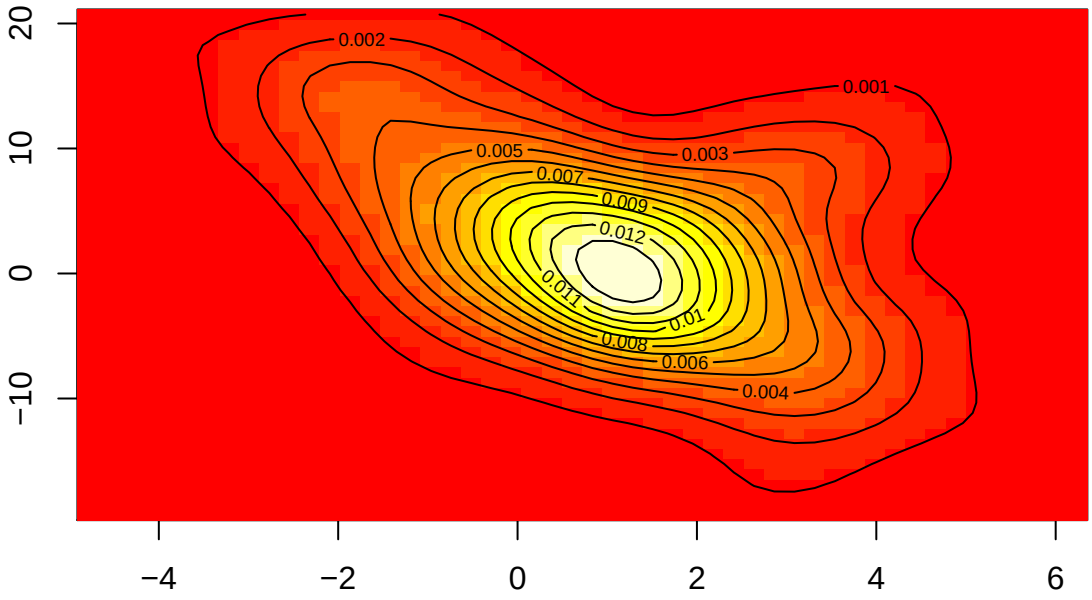
Simulation 2

```
N <- 200 # Number of random samples
set.seed(123)
# Target parameters for univariate normal distributions
```

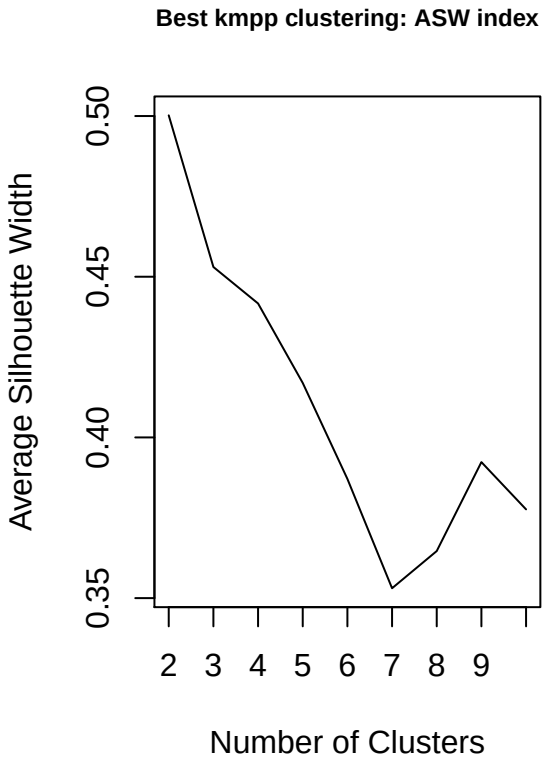
```r
mu1 <- 1; s1 <- 2
mu2 <- 1; s2 <- 8

# Parameters for bivariate normal distribution
mu <- c(mu1,mu2) # Mean
sigma <- list()
rho <- -0.8
sigma[[1]] <- matrix(c(s1^2, s1*s2*rho, s1*s2*rho, s2^2), 2) # Covariance matrix
rho <- 0.8
sigma[[2]] <- matrix(c(s1^2, s1*s2*rho, s1*s2*rho, s2^2), 2) # Covariance matrix

# Function to draw ellipse for bivariate normal data
ellipse_bvn <- function(bvn, alpha){
  Xbar <- apply(bvn,2,mean)
  S <- cov(bvn)
  ellipse(Xbar, S, alpha = alpha, col="red")
}

components <- sample(1:2,prob=c(0.75,0.25),size=N,replace=TRUE)

mydata <- matrix(ncol = 2, nrow = N)
for(i in 1:N) {
  mydata[i, ] <- mvrnorm(1, mu = mu, Sigma = sigma[[components[i]]])
}

plot(mydata,xlab="X1",ylab="X2", col = components + 2, pch = 19)

rho <- -0.8
mu1 <- 1; s1 <- 2
mu2 <- 1; s2 <- 8

# Parameters for bivariate normal distribution
mu <- c(mu1,mu2) # Mean
sigma <- matrix(c(s1^2, s1*s2*rho, s1*s2*rho, s2^2), 2) # Covariance matrix
bvn1 <- mvrnorm(N, mu = mu, Sigma = sigma) # from MASS package

rho <- 0.8
sigma <- matrix(c(s1^2, s1*s2*rho, s1*s2*rho, s2^2), 2) # Covariance matrix
bvn2 <- mvrnorm(N, mu = mu, Sigma = sigma) # from MASS package

ellipse_bvn(bvn1, .05)
ellipse_bvn(bvn2, .05)
```

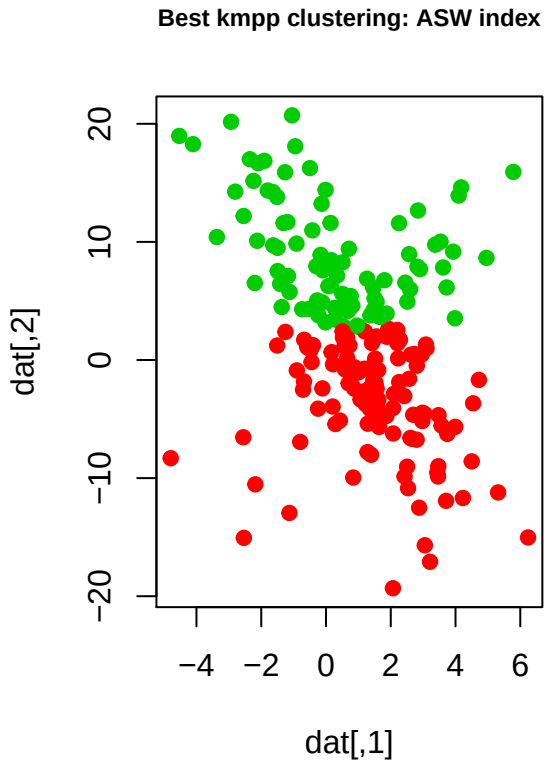


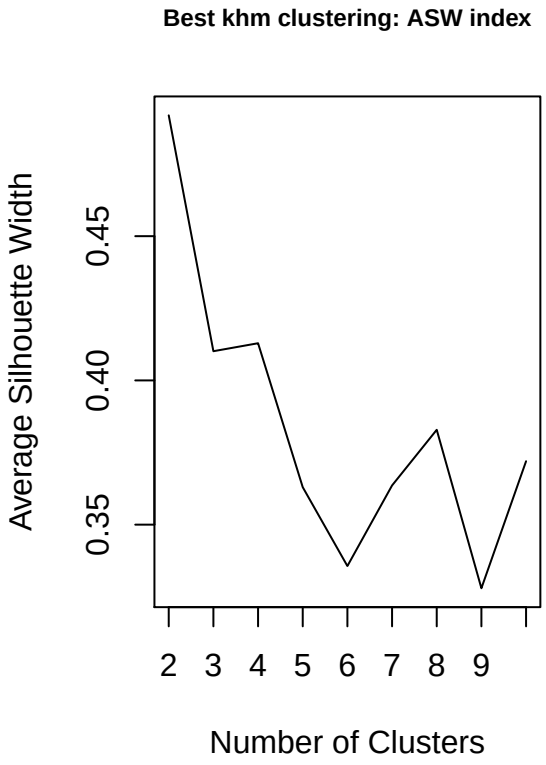

```
bivn.kde <- kde2d(mydata[,1], mydata[,2], n = 50)
image(bivn.kde)       # from base graphics package
contour(bivn.kde, add = TRUE)
```

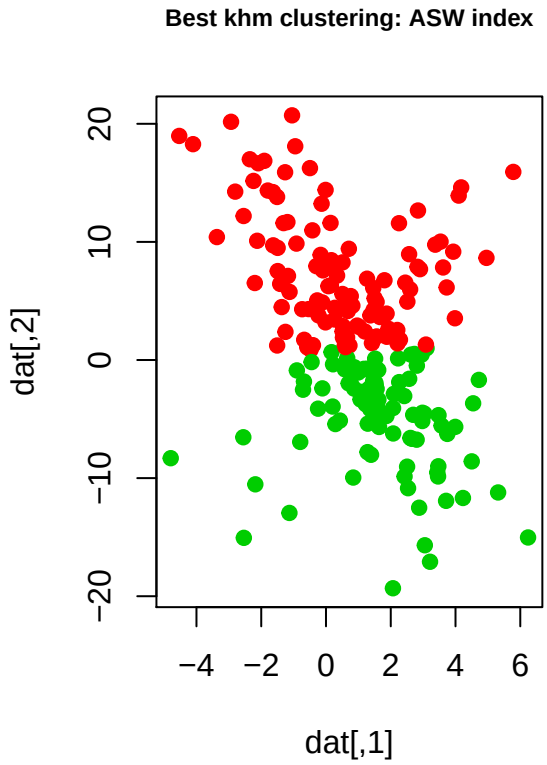


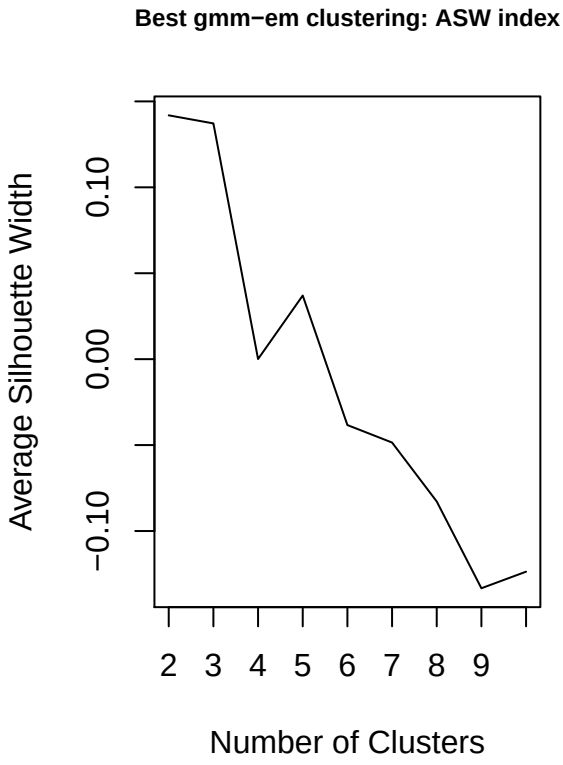

```
truth <- components
df <- simulate(mydata, truth, run.seed = 123, imax = 100, ks = seq(2, 10, 1), cov_model = "EEV")
```

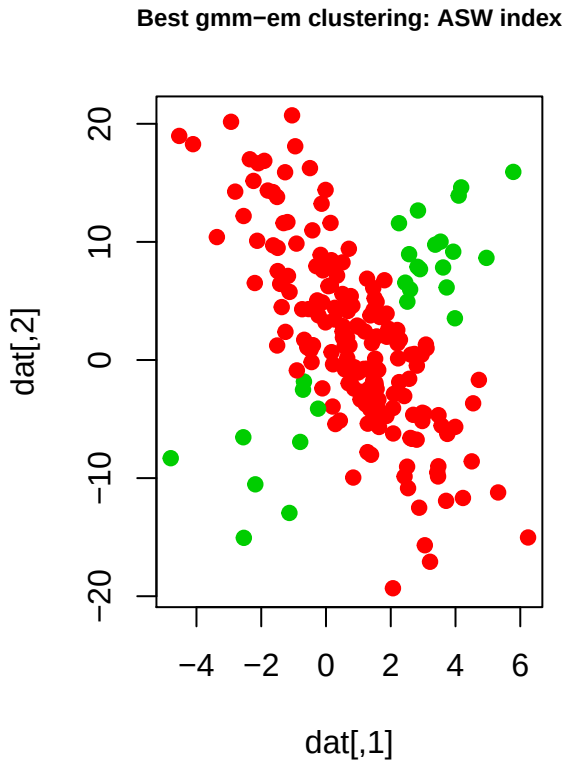

Best kmpp clustering: ASW index
Average Silhouette Width
0.35
0.40
0.45
0.50
2 3 4 5 6 7 8 9
Number of Clusters

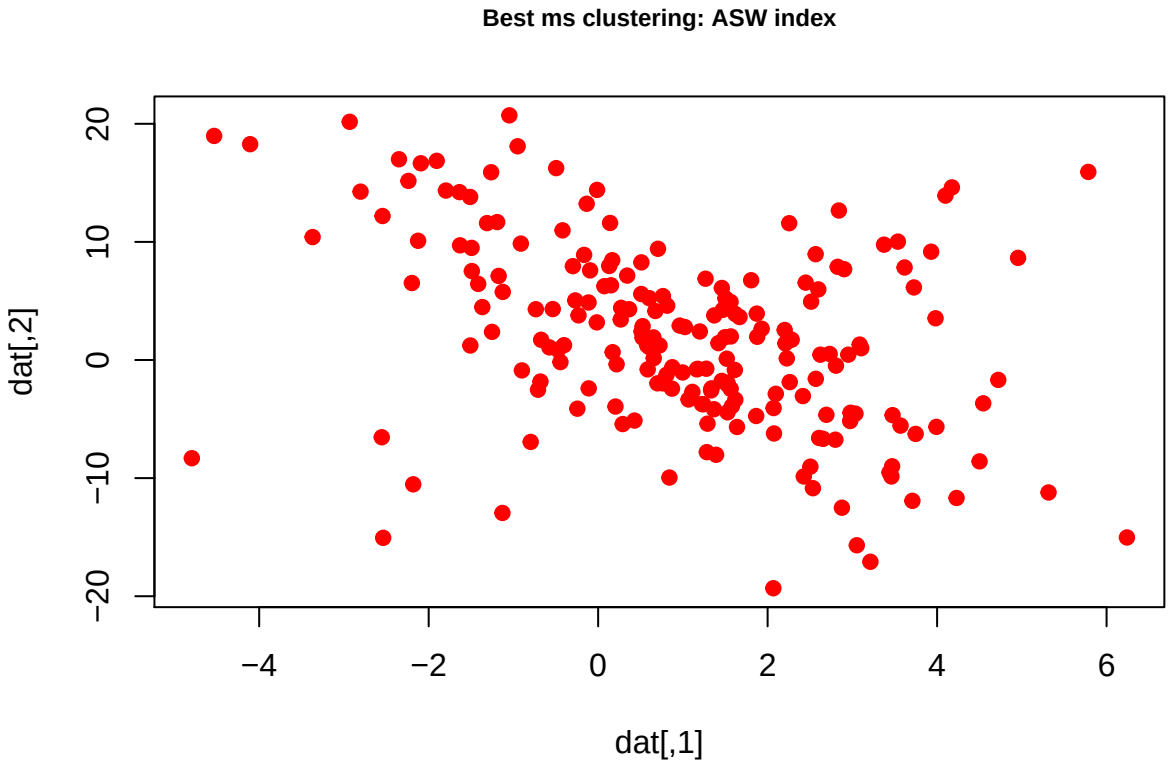

Best kmpp clustering: ASW index
dat[,2]
−20
−10
0
10
20
−4 −2 0 2 4 6
dat[,1]

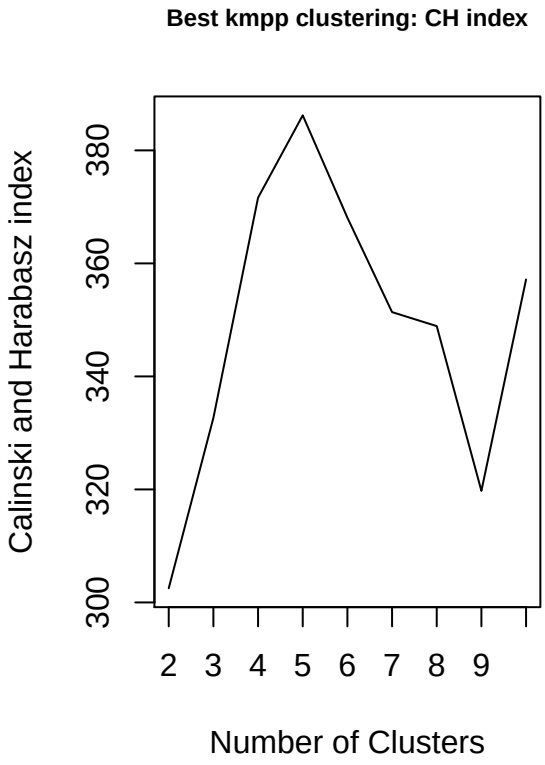

Best khm clustering: ASW index
Average Silhouette Width
0.35
0.40
0.45
2 3 4 5 6 7 8 9
Number of Clusters

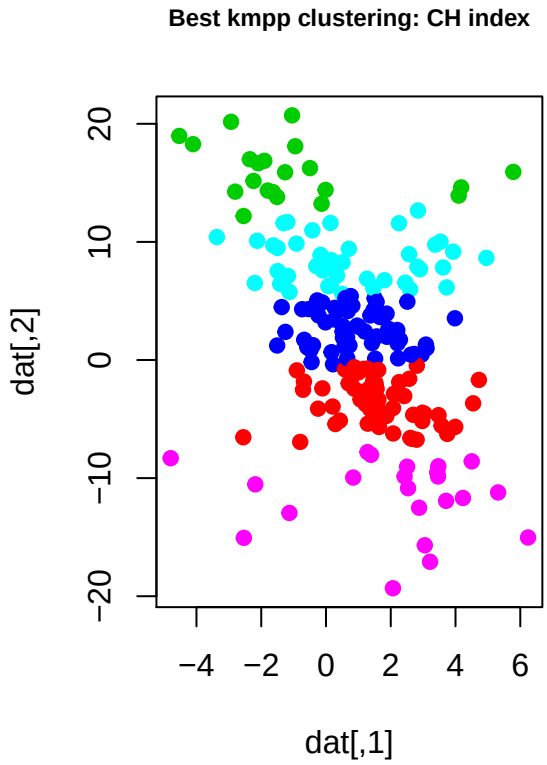

Best khm clustering: ASW index
dat[,2]
−20
−10
0
10
20
−4 −2 0 2 4 6
dat[,1]

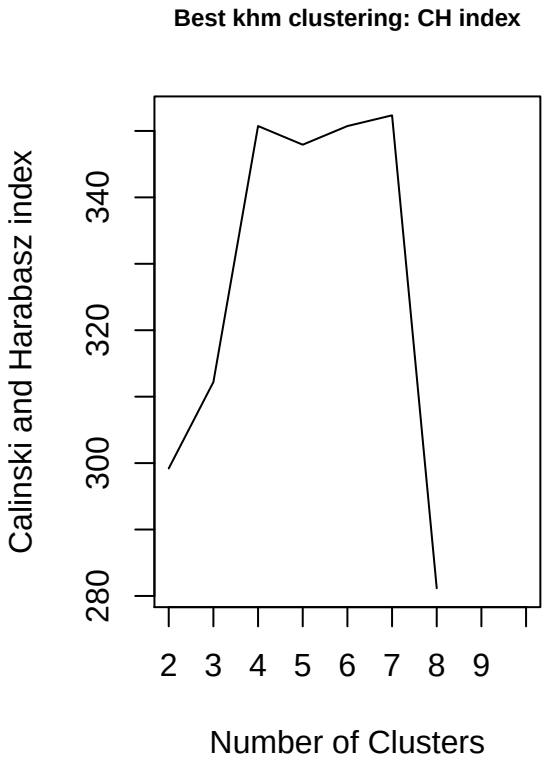

Best gmm−em clustering: ASW index
Average Silhouette Width
−0.10
0.00
0.10
2 3 4 5 6 7 8 9
Number of Clusters

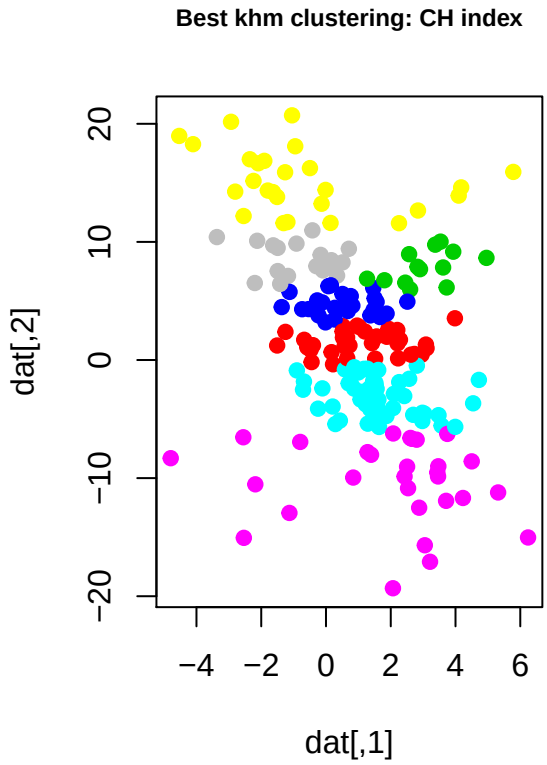

Best gmm−em clustering: ASW index
dat[,2]
−20
−10
0
10
20
−4 −2 0 2 4 6
dat[,1]

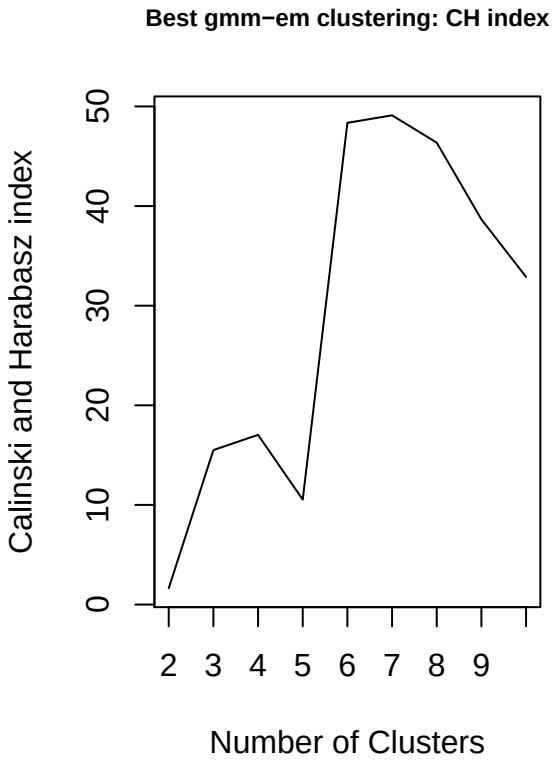

Best ms clustering: ASW index
dat[,2]
20
10
0
−10
−20
−4
−2
0
2
4
6
dat[,1]

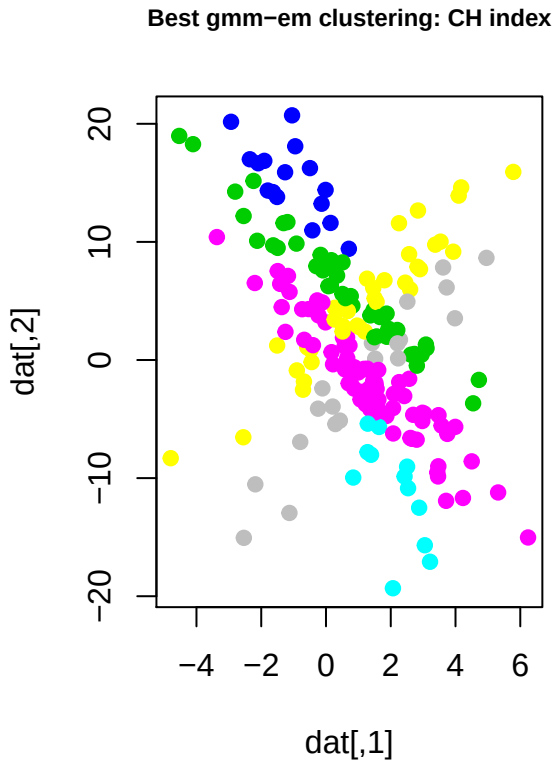

Best kmpp clustering: CH index
Calinski and Harabasz index
300
320
340
360
380
2
3
4
5
6
7
8
9
Number of Clusters


Best kmpp clustering: CH index
dat[,2]
−20
−10
0
10
20
−4
−2
0
2
4
6
dat[,1]


Best khm clustering: CH index
Calinski and Harabasz index
280
300
320
340
2
3
4
5
6
7
8
9
Number of Clusters


Best khm clustering: CH index
dat[,2]
−20
−10
0
10
20
−4
−2
0
2
4
6
dat[,1]


Best gmm−em clustering: CH index
Calinski and Harabasz index
0
10
20
30
40
50
2
3
4
5
6
7
8
9
Number of Clusters


Best gmm−em clustering: CH index
dat[,2]
−20
−10
0
10
20
−4
−2
0
2
4
6
dat[,1]

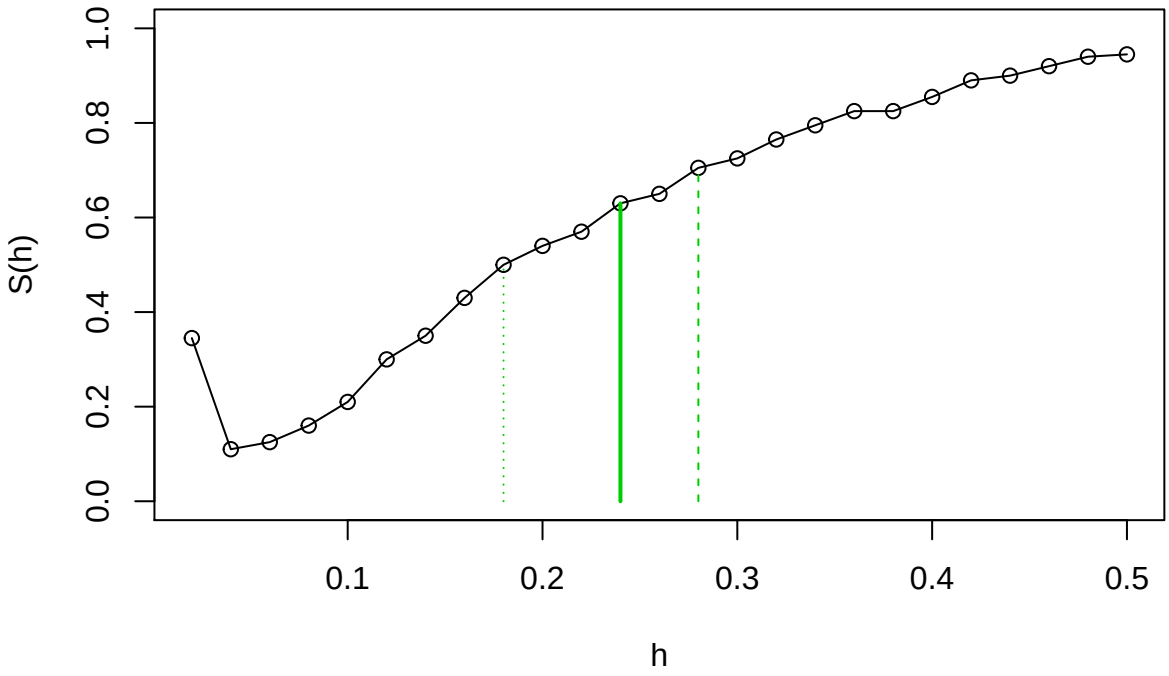
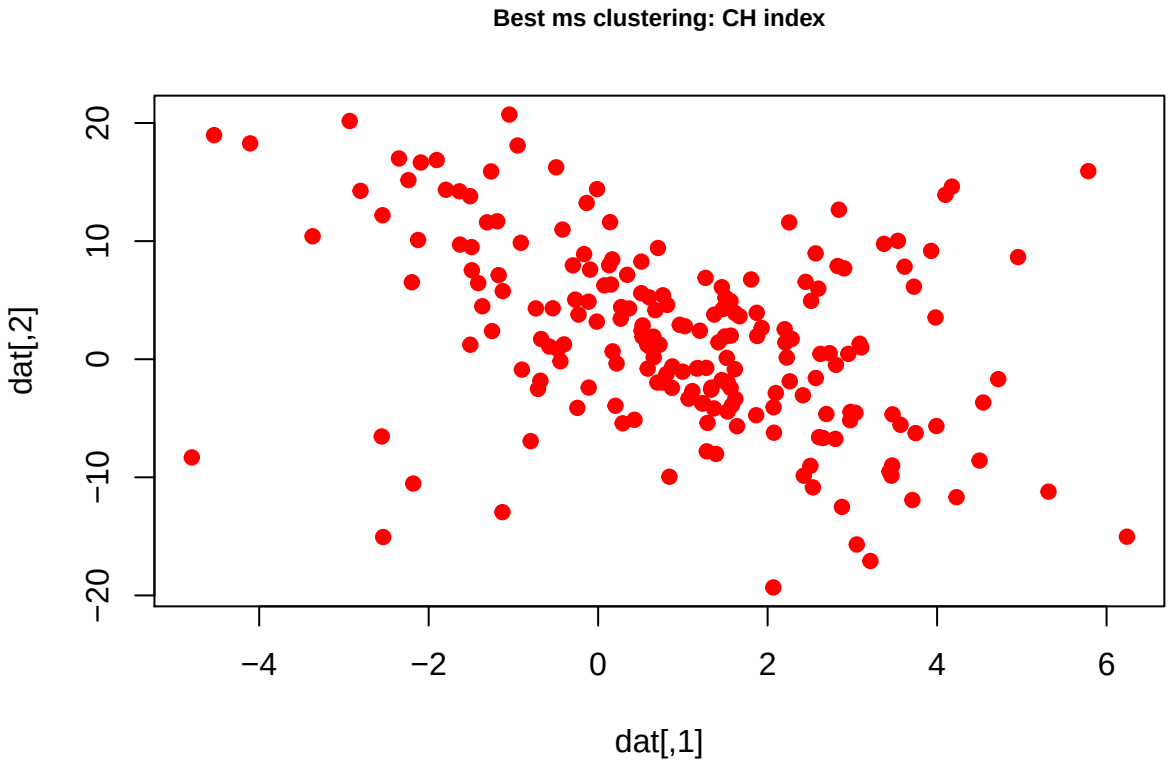

S(h)
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.2
0.3
0.4
0.5
h

Table 2: External validation measures for simulation data set 2. Same caption as Table 1

	Average Silhouette Width							*Calinski and Harabasz index*						
	RI	HA	MA	FM	JI	VI	NMI	RI	HA	MA	FM	JI	VI	NMI
k-means++	0.50	0.00	0.00	0.57	0.39	1.77	0.00	0.42	-0.01	0.00	0.38	0.20	2.96	0.01
k-harmonic means	0.50	0.00	0.00	0.56	0.39	1.78	0.00	0.42	0.02	0.02	0.34	0.16	3.25	0.07
GMM-EM	**0.80**	**0.54**	**0.54**	**0.86**	**0.75**	**0.77**	**0.43**	**0.49**	**0.11**	**0.11**	**0.47**	**0.26**	**2.60**	**0.16**
Mean Shift	0.64	0.00	0.00	0.80	0.64	0.79	0.00	0.64	0.00	0.00	0.80	0.64	0.79	0.00



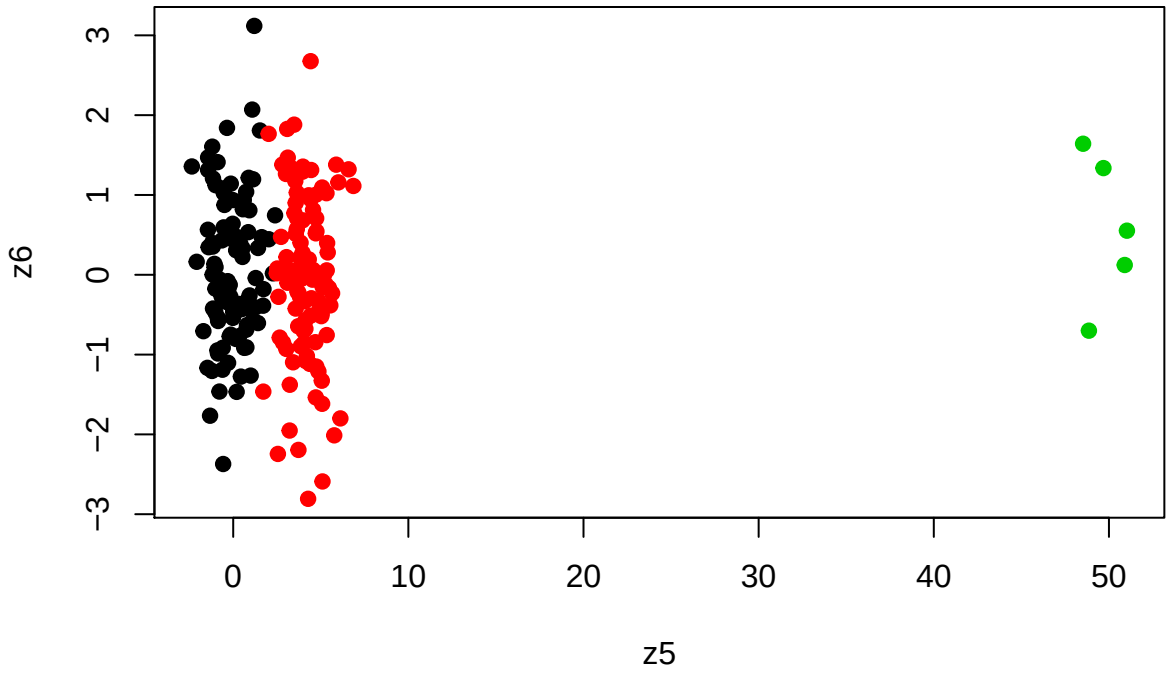

```
dat2 <- mydata
truth2 <- truth
```

```
df <- round(df, 2)
rownames(df) <- c("k-means++", "k-harmonic means", "GMM-EM", "Mean Shift")

kable(df, format = "latex", align = "c", booktabs = T, caption = "External validation measures for simul
row_spec(3, bold = T)
```

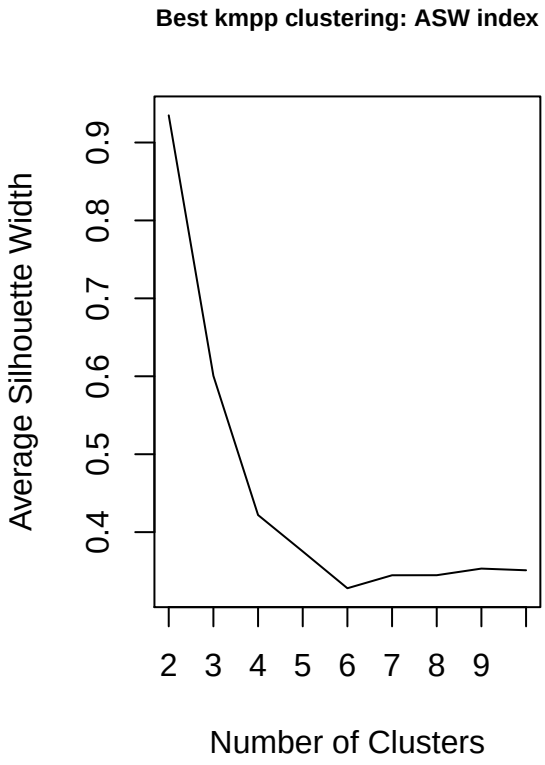
Simulation 3

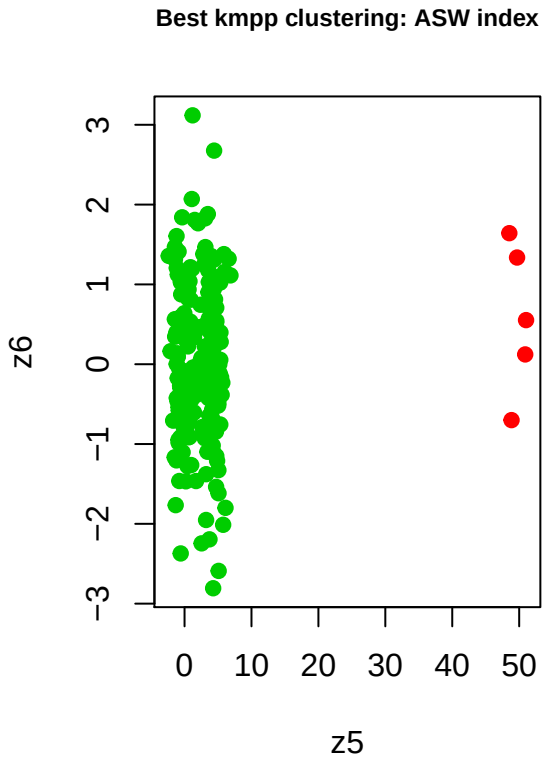
```
set.seed(11118)
z1 <- rnorm(100,0,1)
z2 <- rnorm(100,4,1)
z3 <- rnorm(100,0,1)
```

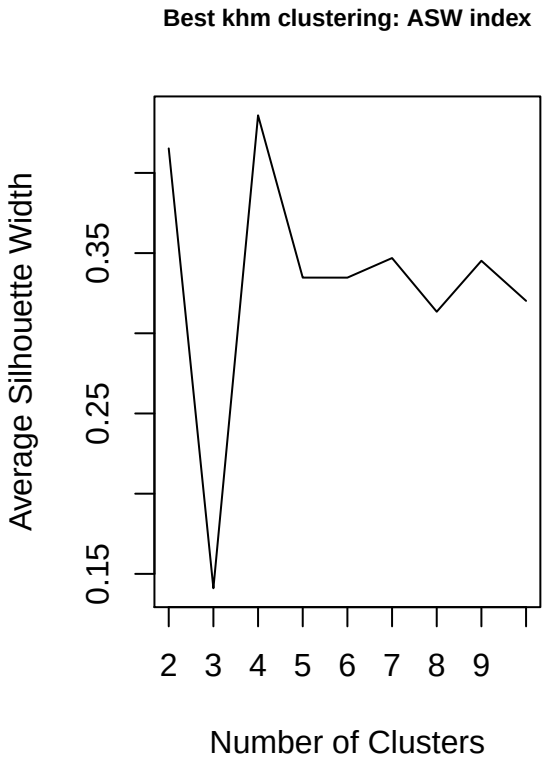
```r
z4 <- rnorm(100,0,1)
z5 <- rnorm(5,50,1)
z6 <- rnorm(5,0,1)
# z7 <- rnorm(5,50,1)
# z8 <- rnorm(5,-4,1)
za <- cbind(c(z1,z2),c(z3,z4))
zb <- rbind(za, cbind(z5, z6))
truth <- c(rep(1, 100), rep(2, 100), rep(3, 5))
plot(zb, col = truth, pch=19)
```

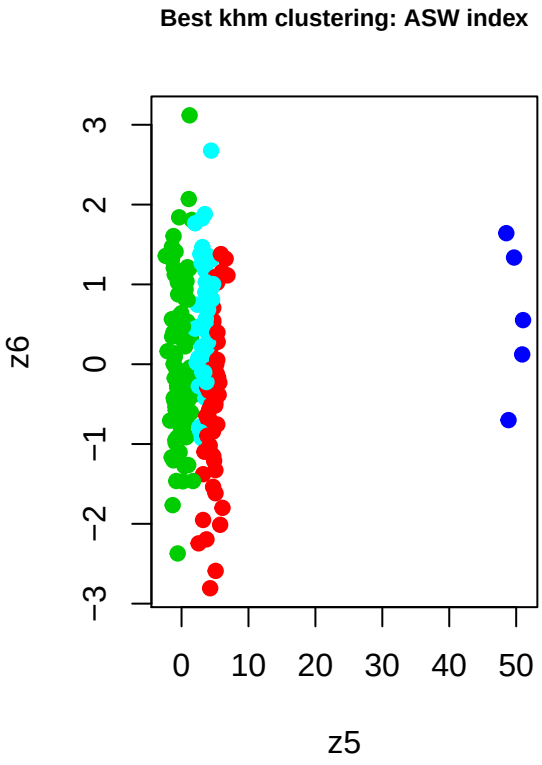


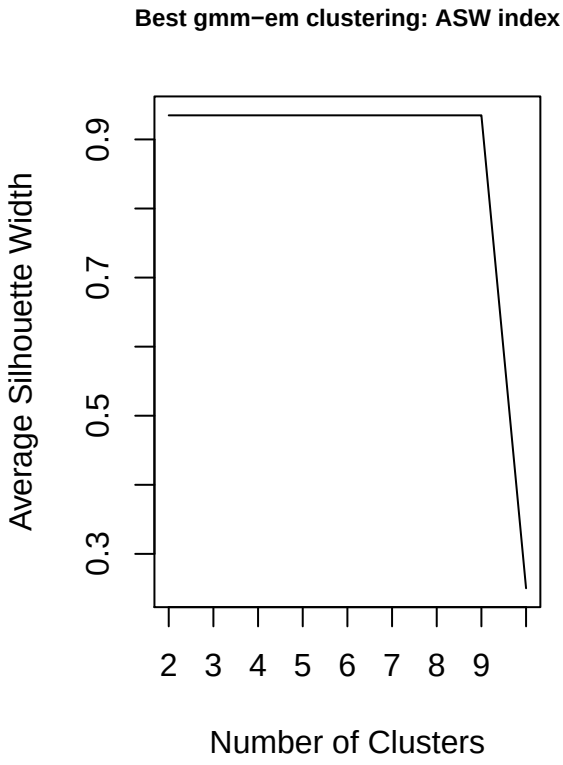

```r
mydata <- zb

df <- simulate(mydata, truth, run.seed = 154, imax = 100, ks = seq(2, 10, 1))
```

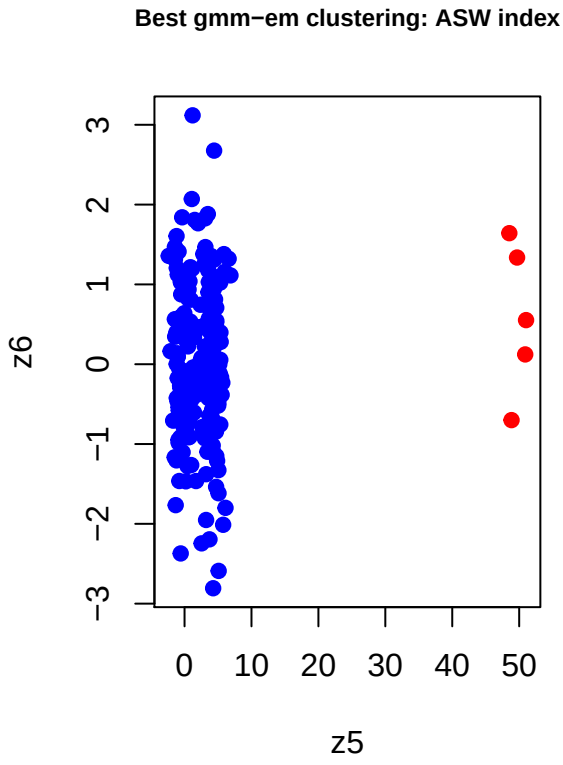

Best kmpp clustering: ASW index
Average Silhouette Width
0.4
0.5
0.6
0.7
0.8
0.9
2
3
4
5
6
7
8
9
Number of Clusters


Best kmpp clustering: ASW index
z6
−3
−2
−1
0
1
2
3
0
10
20
30
40
50
z5

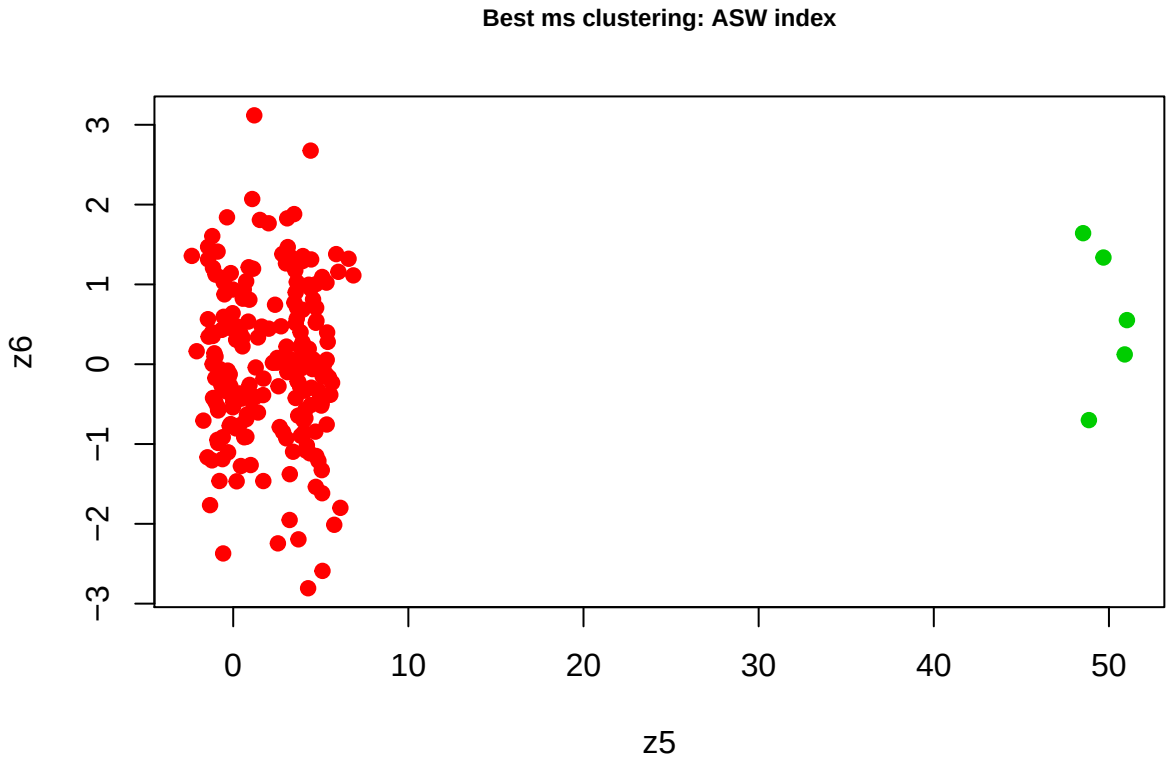

Best khm clustering: ASW index
Average Silhouette Width
0.15
0.25
0.35
2 3 4 5 6 7 8 9
Number of Clusters

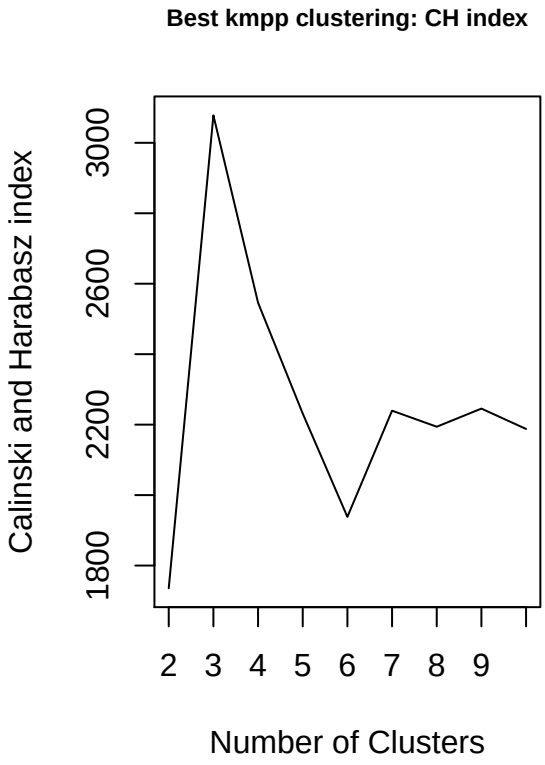

Best khm clustering: ASW index
z6
−3 −2 −1 0 1 2 3
0 10 20 30 40 50
z5

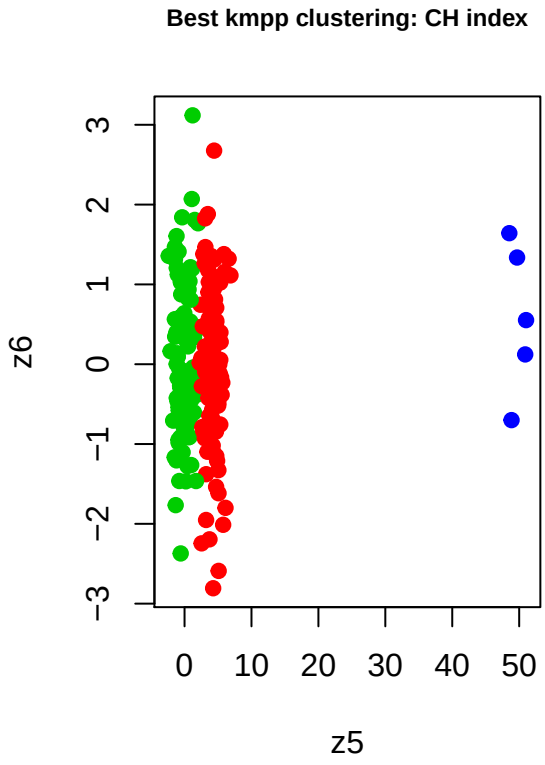


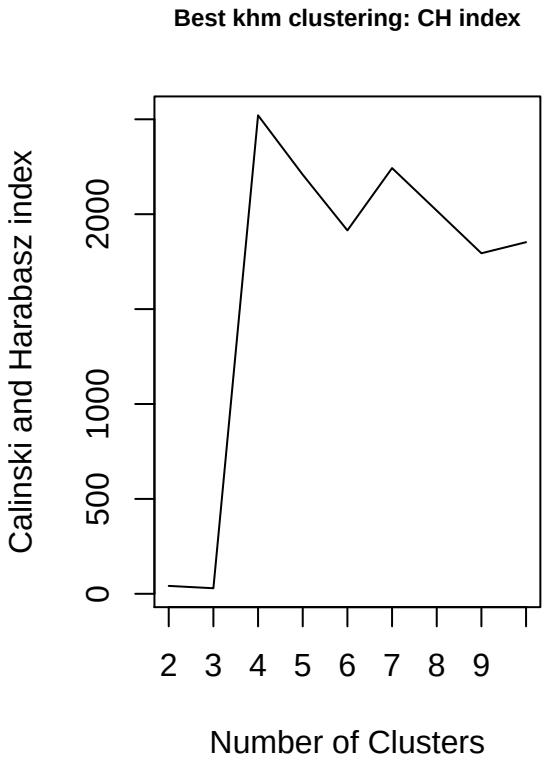


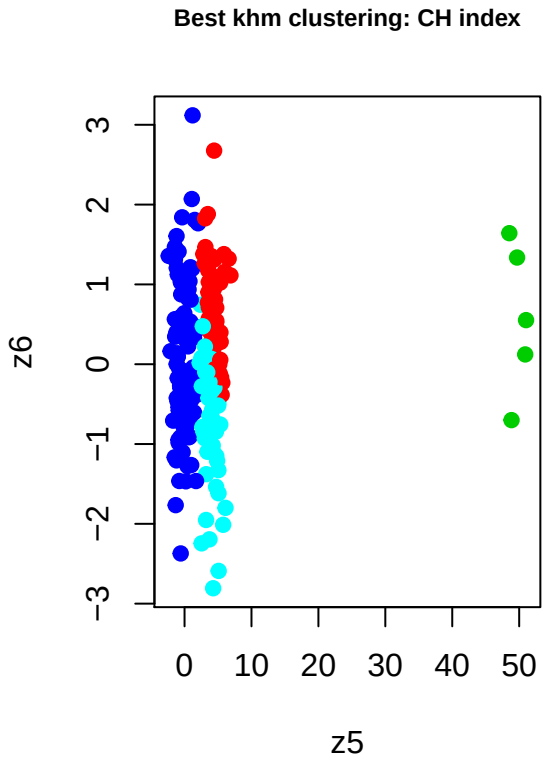

```
## [1] "Selection of the bandwidth parameter"
```

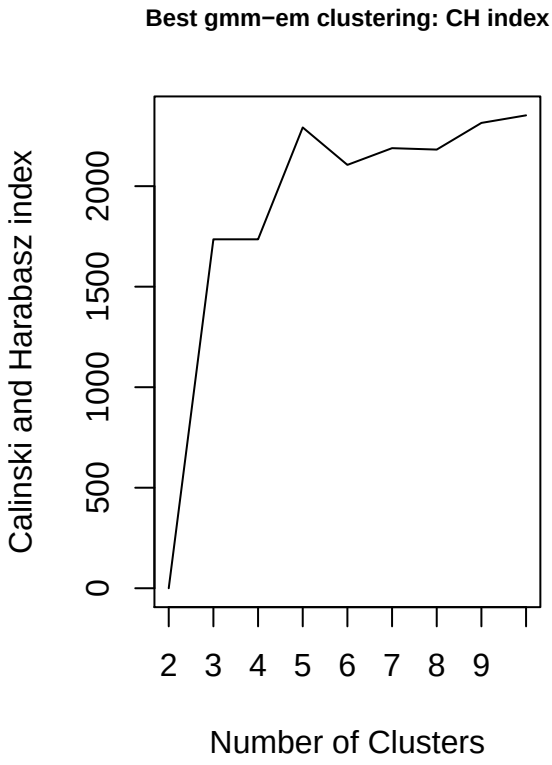

Best ms clustering: ASW index
z6
3
2
1
0
−1
−2
−3
0
10
20
30
40
50
z5

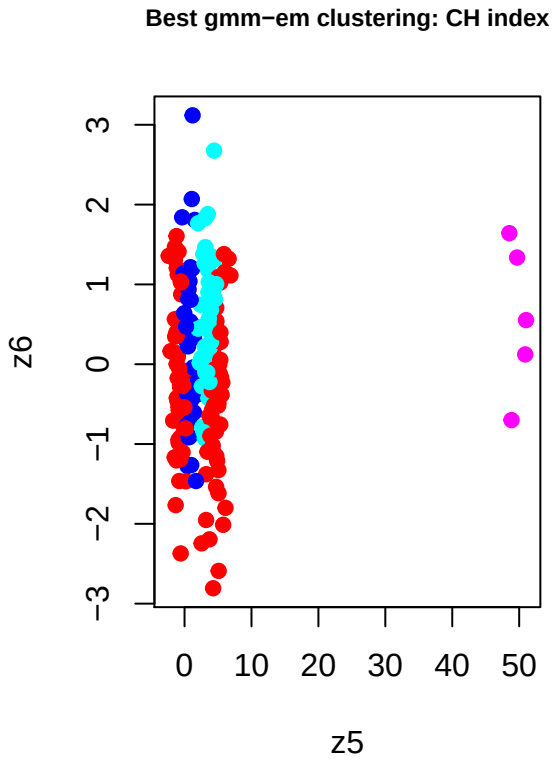

Best kmpp clustering: CH index
Calinski and Harabasz index
1800
2200
2600
3000
2 3 4 5 6 7 8 9
Number of Clusters

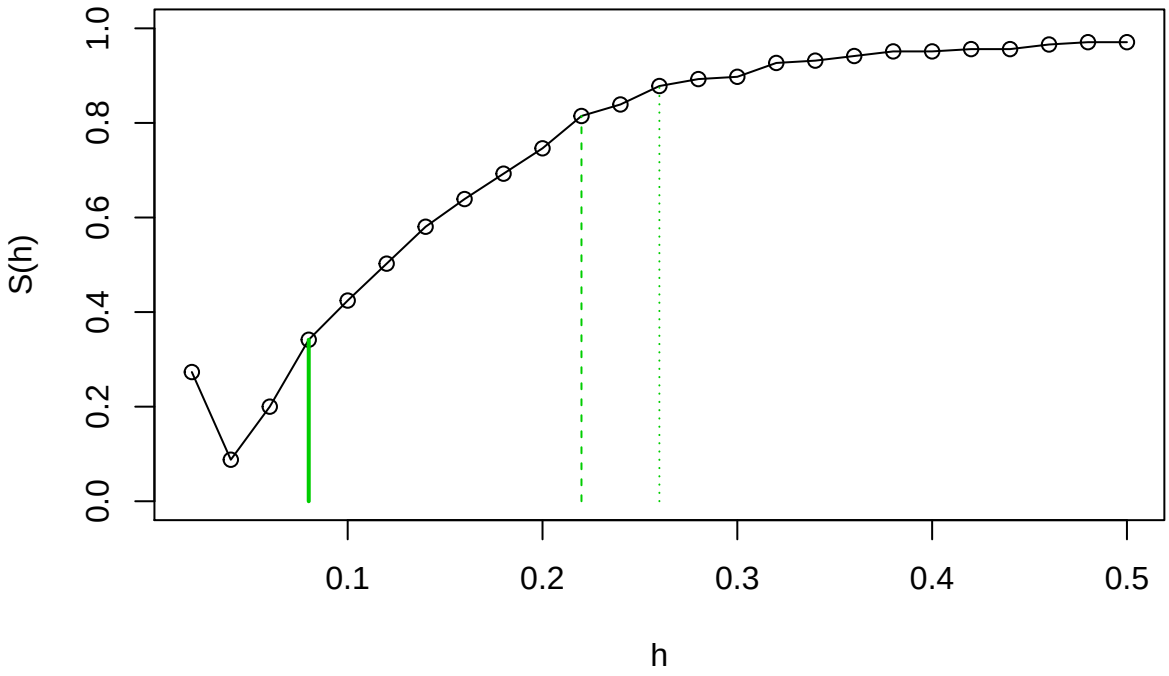

Best kmpp clustering: CH index
z6
−3 −2 −1 0 1 2 3
0 10 20 30 40 50
z5


Best khm clustering: CH index
Calinski and Harabasz index
0
500
1000
2000
2 3 4 5 6 7 8 9
Number of Clusters


Best khm clustering: CH index
z6
−3 −2 −1 0 1 2 3
0 10 20 30 40 50
z5


Best gmm−em clustering: CH index
Calinski and Harabasz index
0
500
1000
1500
2000
2 3 4 5 6 7 8 9
Number of Clusters

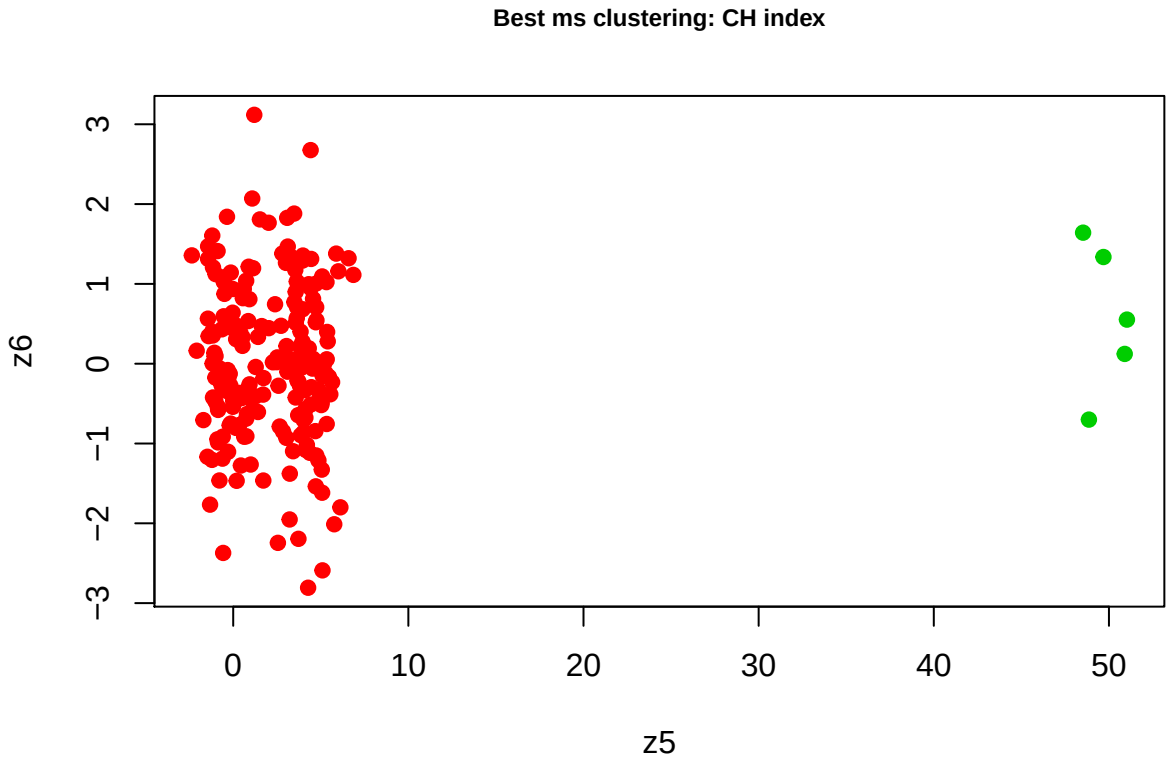

Best gmm−em clustering: CH index
z6
−3 −2 −1 0 1 2 3
0 10 20 30 40 50
z5




```
## [1] "Selection of the bandwidth parameter"
```

Table 3: External validation measures for simulation data set 3. Same caption as Table 1.

	Average Silhouette Width							*Calinski and Harabasz index*						
	RI	HA	MA	FM	JI	VI	NMI	RI	HA	MA	FM	JI	VI	NMI
k-means++	**0.52**	**0.09**	**0.09**	**0.71**	**0.5**	**0.98**	**0.25**	**0.96**	**0.92**	**0.93**	**0.96**	**0.92**	**0.28**	**0.88**
k-harmonic means	0.85	0.70	0.71	0.83	0.7	0.73	0.74	0.85	0.70	0.70	0.83	0.70	0.74	0.73
GMM-EM	0.52	0.09	0.09	0.71	0.5	0.98	0.25	0.75	0.48	0.49	0.68	0.48	1.19	0.63
Mean Shift	0.52	0.09	0.09	0.71	0.5	0.98	0.25	0.52	0.09	0.09	0.71	0.50	0.98	0.25



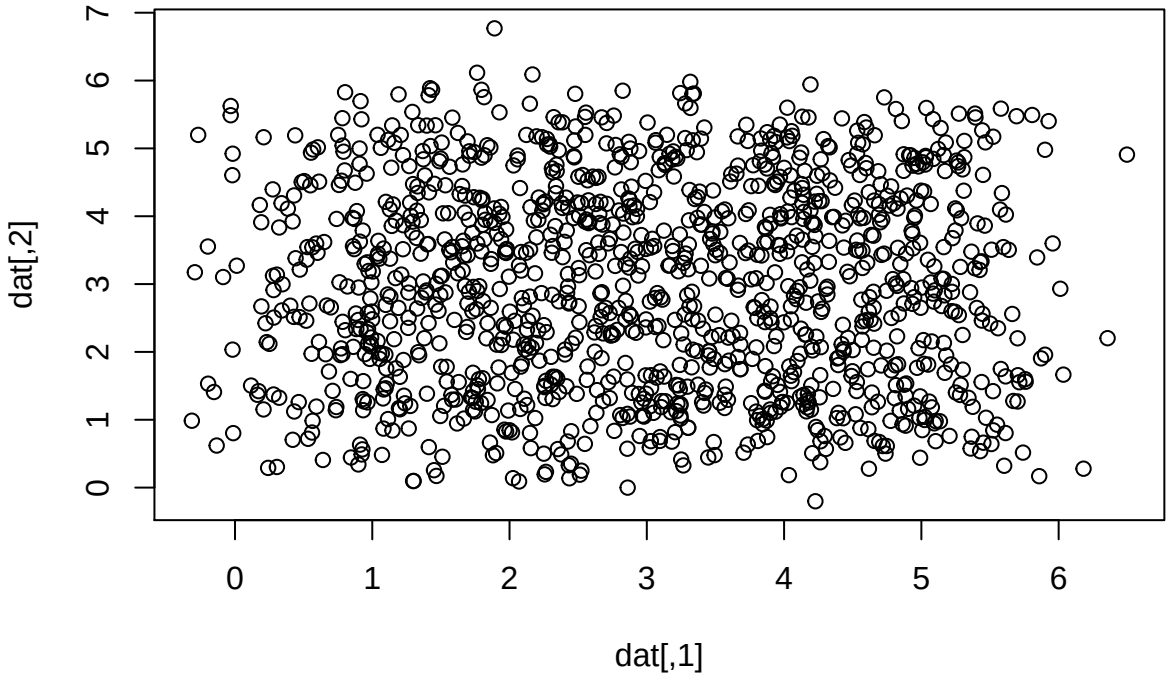

```
dat3 <- mydata
truth3 <- truth
```

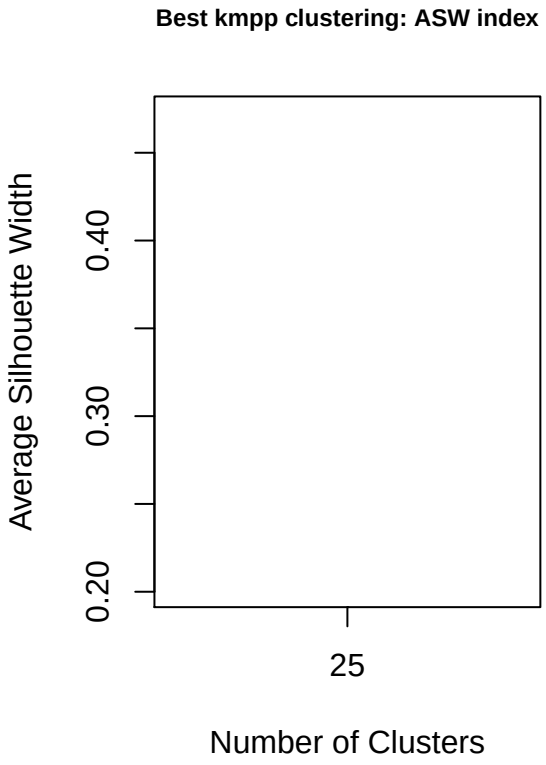
```
df <- round(df, 2)
rownames(df) <- c("k-means++", "k-harmonic means", "GMM-EM", "Mean Shift")

kable(df, format = "latex", align = "c", booktabs = T, caption = "External validation measures for simul
row_spec(1, bold = T)
```

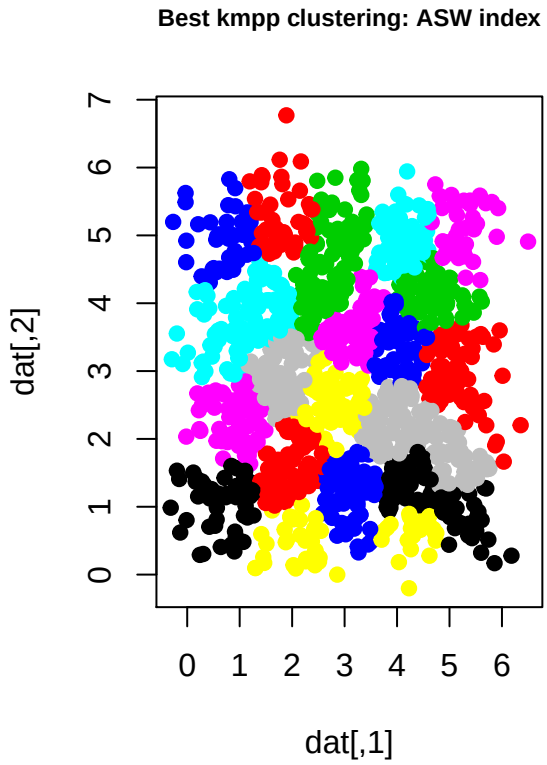
Simultation 4

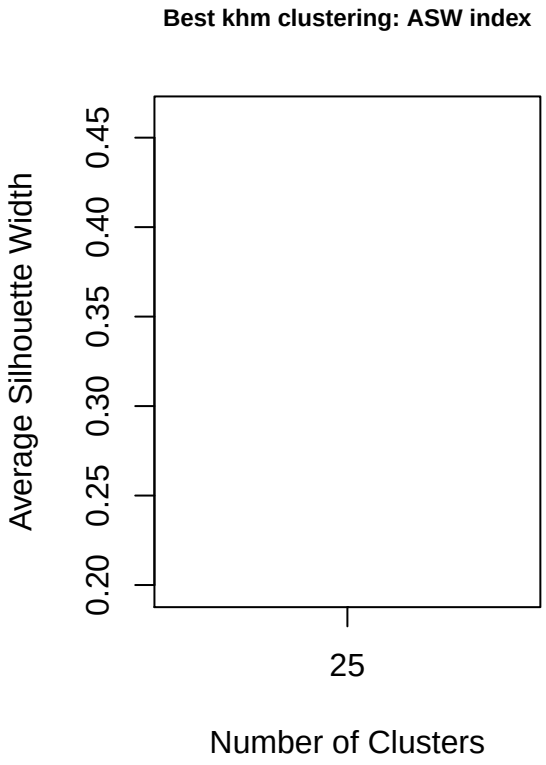
```
simulation_results <- list()
param_results <- list()
grid <- c(5, 10)
for(z in 1) {
```

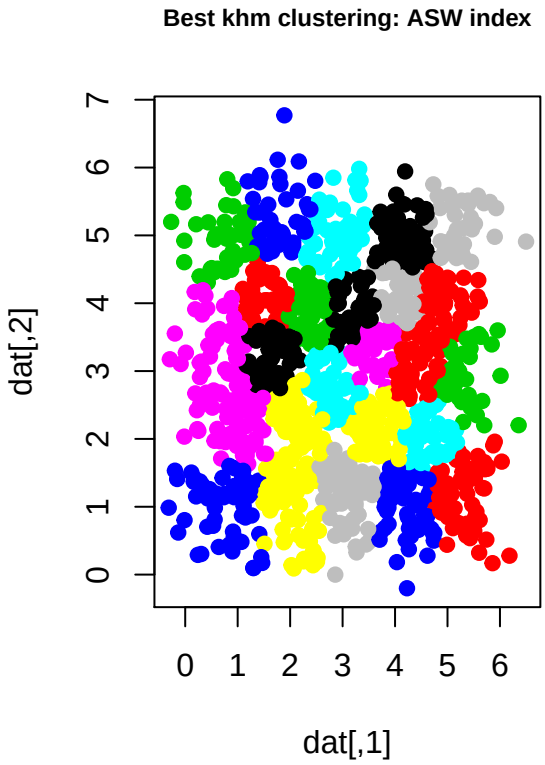
```
  l <- grid[z]
  k <- grid[z]
  for(m in c(50)) {
    mu <- matrix(ncol = 2, nrow = l*k)
    idx <- 1
    for(i in 1:l) {
      for(j in 1:k) {
        mu[idx, ] <- c(i, j)
        idx <- idx + 1
      }
    }

    dat <- matrix(ncol = 2, nrow = l*k*m)

    for(i in 1:nrow(mu)) {
      dat[1:m + m*(i-1), 1] <- rnorm(m, mu[i, 1], .5)
      dat[1:m + m*(i-1), 2] <- rnorm(m, mu[i, 2], .5)
    }
    plot(dat)

    truth <- vector(length = l*k*m)
    for(i in 1:nrow(mu)) {
      truth[1:m + m*(i-1)] <- rep(i, m)
    }

    df <- simulate(dat, truth, run.seed = 154, imax = 100, ks = l*k)

    dat4 <- dat
    truth4 <- truth
  }
}
```

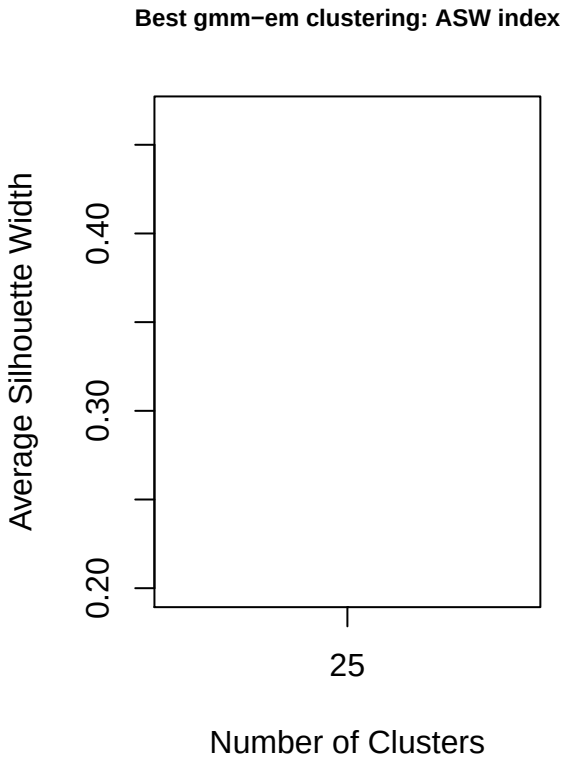

dat[,2]
0
1
2
3
4
5
6
7
dat[,1]
0
1
2
3
4
5
6

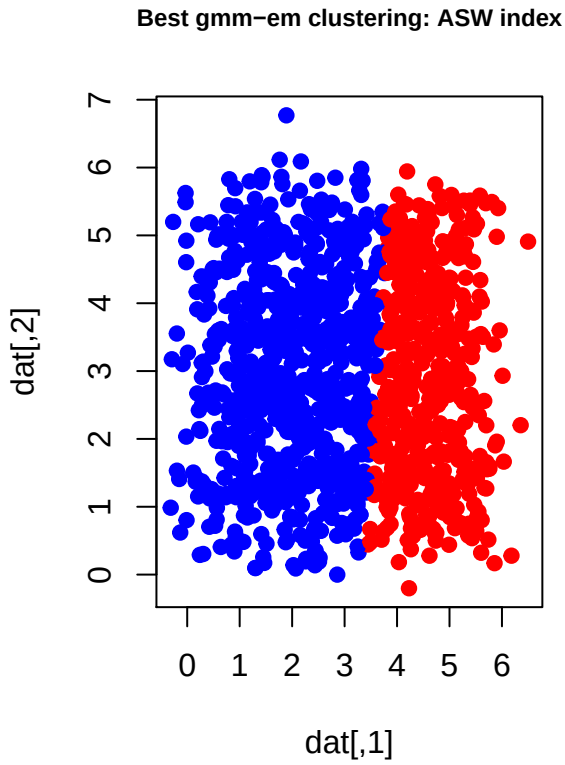

Best kmpp clustering: ASW index
Average Silhouette Width
0.20
0.30
0.40
25
Number of Clusters


Best kmpp clustering: ASW index
dat[,2]
0
1
2
3
4
5
6
7
dat[,1]

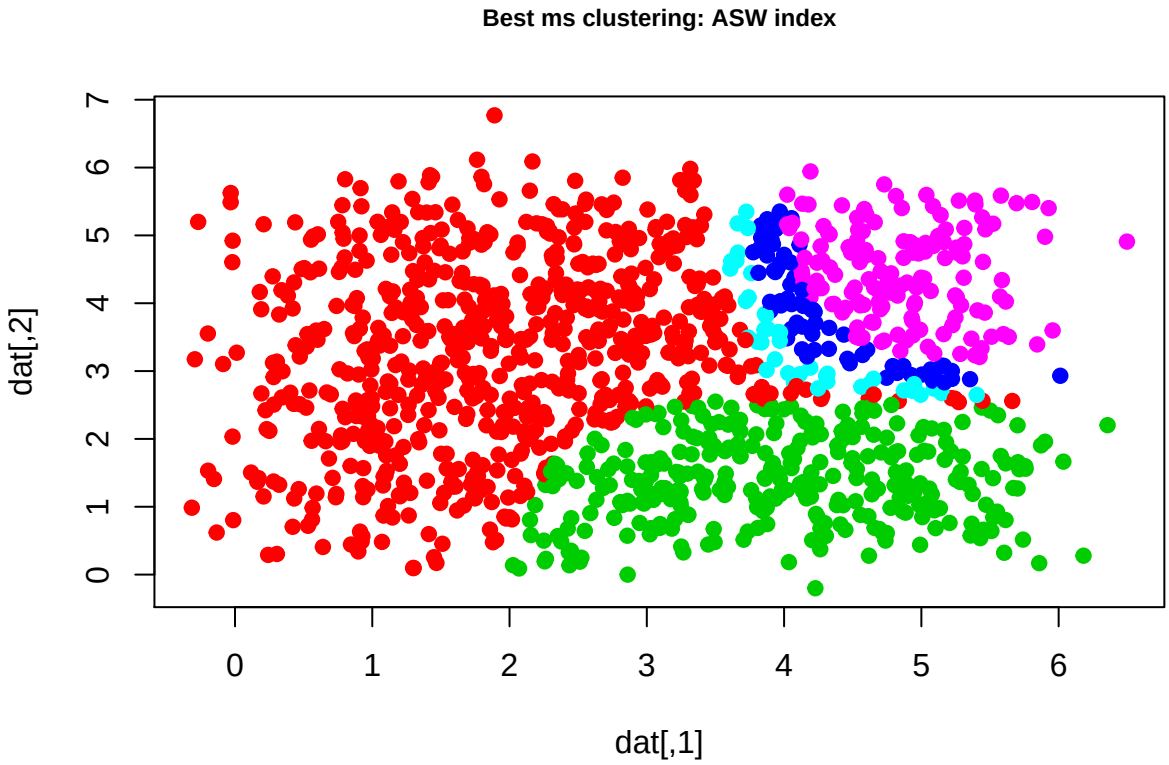

Best khm clustering: ASW index
Average Silhouette Width
0.20
0.25
0.30
0.35
0.40
0.45
25
Number of Clusters

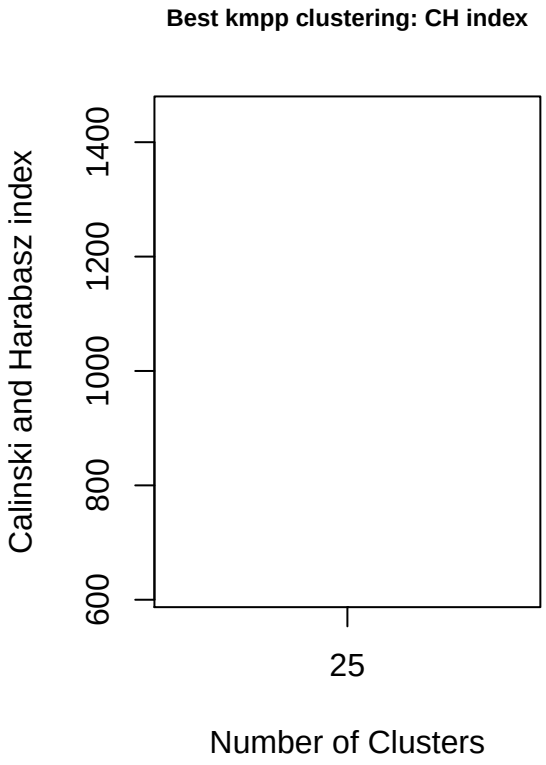

Best khm clustering: ASW index
dat[,2]
0
1
2
3
4
5
6
7
dat[,1]

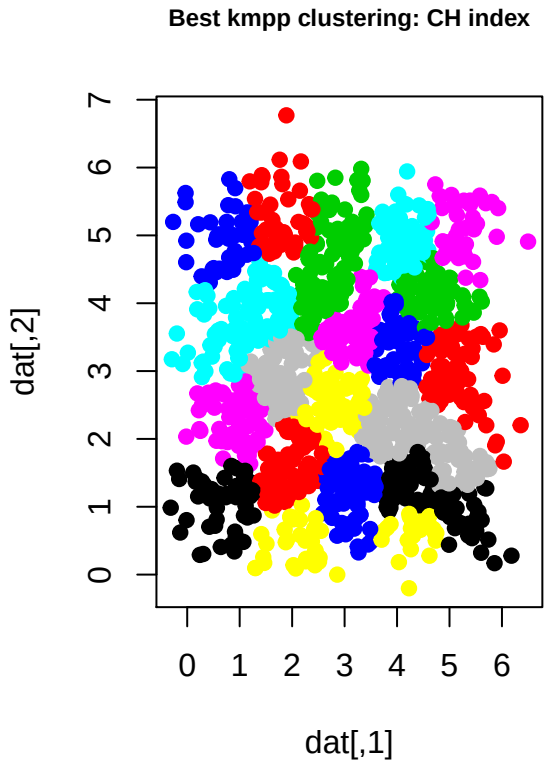


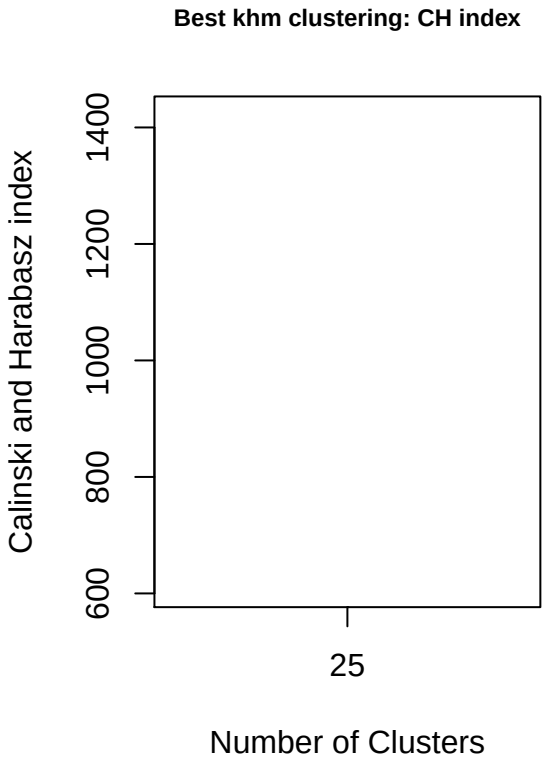


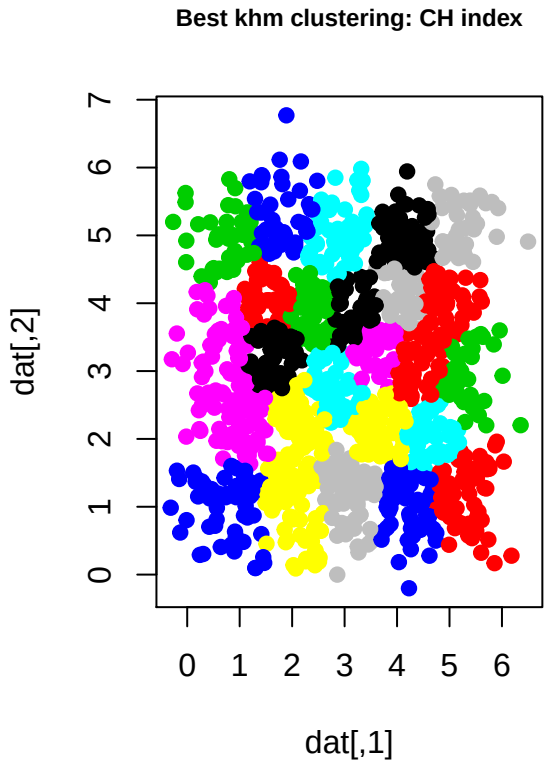

```
## [1] "Selection of the bandwidth parameter"
```

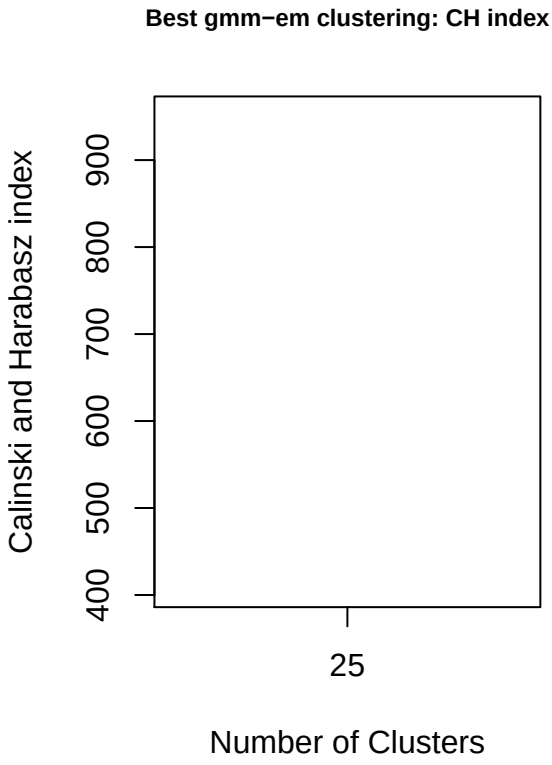

Best ms clustering: ASW index
dat[,2]
dat[,1]
0
1
2
3
4
5
6
7

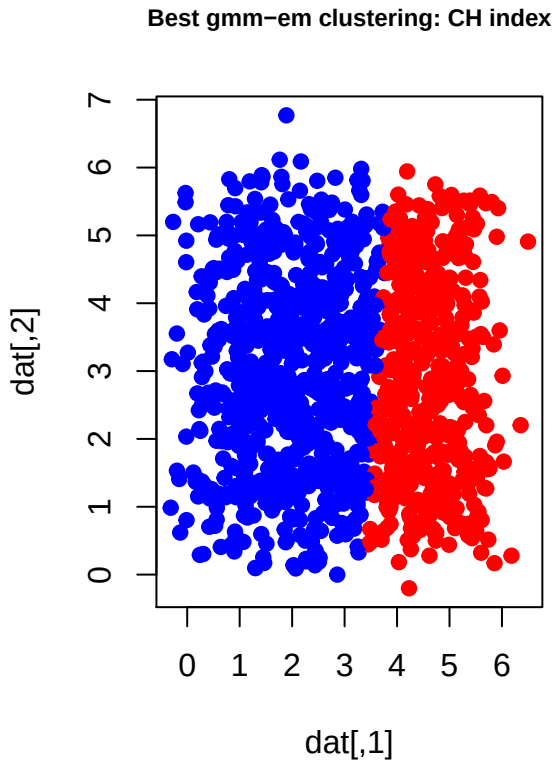

Best kmpp clustering: CH index
Calinski and Harabasz index
600
800
1000
1200
1400
25
Number of Clusters

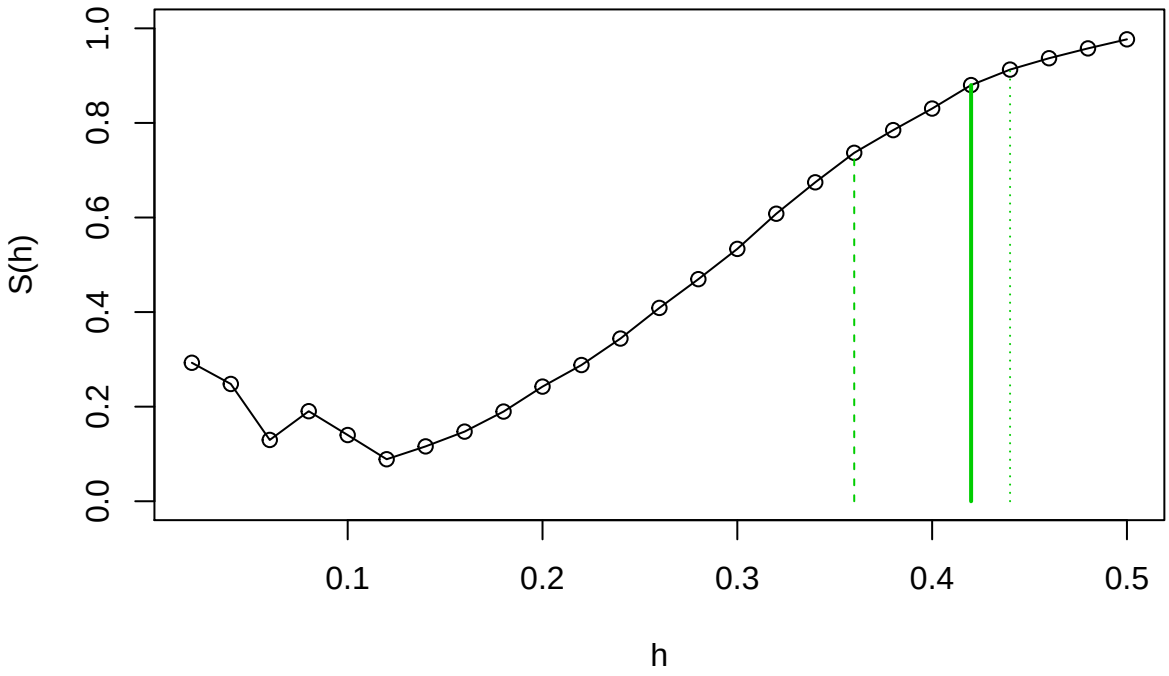

Best kmpp clustering: CH index
dat[,2]
0
1
2
3
4
5
6
7
0 1 2 3 4 5 6
dat[,1]


Best khm clustering: CH index
Calinski and Harabasz index
600
800
1000
1200
1400
25
Number of Clusters

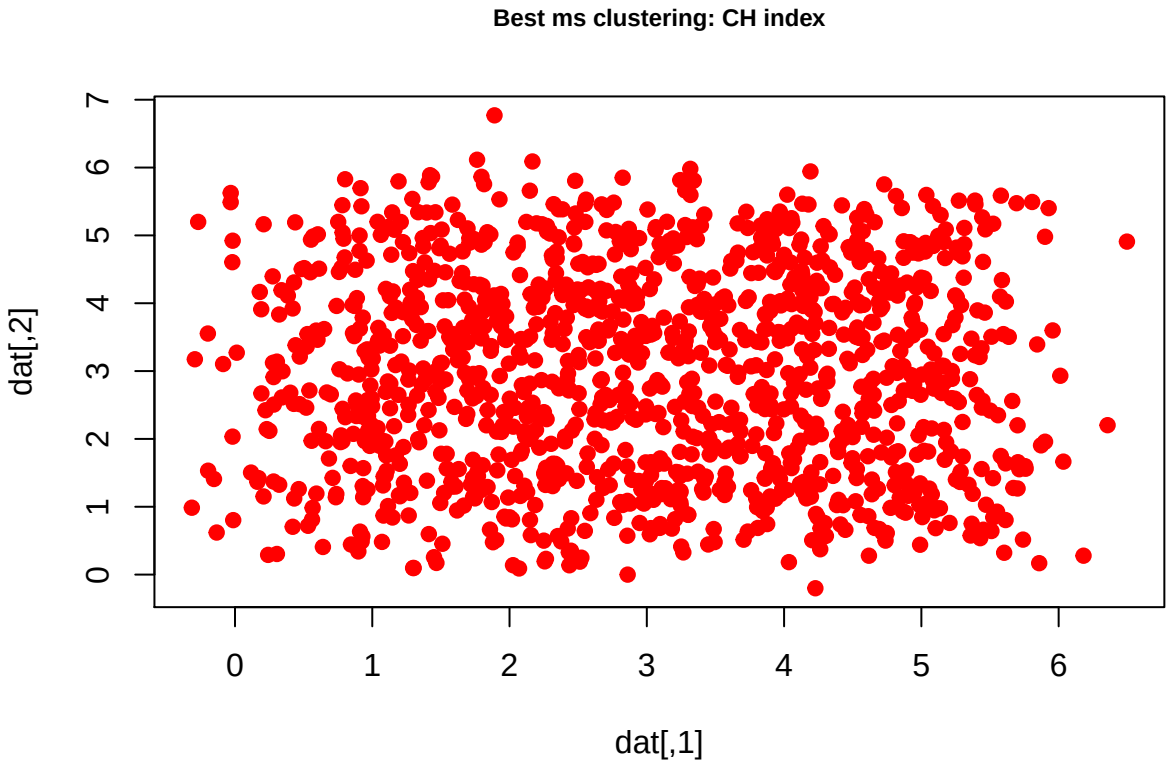

Best khm clustering: CH index
dat[,2]
0
1
2
3
4
5
6
7
0
1
2
3
4
5
6
dat[,1]


Best gmm−em clustering: CH index
Calinski and Harabasz index
400
500
600
700
800
900
25
Number of Clusters


Best gmm−em clustering: CH index
dat[,2]
0
1
2
3
4
5
6
7
0
1
2
3
4
5
6
dat[,1]




```
## [1] "Selection of the bandwidth parameter"
```




```
df
```

```
##             RI         HA         MA        FM         JI       VI
## [1,] 0.9444458 0.28855796 0.30172473 0.3176805 0.18868913 3.880284
## [2,] 0.9468669 0.30739668 0.32045958 0.3351141 0.20124405 3.830526
## [3,] 0.5042575 0.05535948 0.05666674 0.2448092 0.06629269 4.159785
## [4,] 0.6521364 0.08424610 0.08657618 0.2532973 0.08061633 4.067529
##            NMI         RI         HA         MA        FM         JI
## [1,] 0.5795231 0.94444580 0.28855796 0.30172473 0.3176805 0.18868913
## [2,] 0.5864499 0.94686693 0.30739668 0.32045958 0.3351141 0.20124405
## [3,] 0.2577282 0.50425749 0.05535948 0.05666674 0.2448092 0.06629269
## [4,] 0.3636795 0.03923139 0.00000000 0.00000000 0.1980691 0.03923139
##            VI       NMI
## [1,] 3.880284 0.5795231
## [2,] 3.830526 0.5864499
## [3,] 4.159785 0.2577282
## [4,] 4.643856 0.0000000
```

```
df <- round(df, 2)
rownames(df) <- c("k-means++", "k-harmonic means", "GMM-EM", "Mean Shift")

kable(df, format = "latex", align = "c", booktabs = T, caption = "External validation measures for simul
row_spec(2, bold = T)
```

Table 4: External validation measures for simulation data set 4. Same caption as Table 1, except that only the correct k=25 was considered

	Average Silhouette Width							*Calinski and Harabasz index*						
	RI	HA	MA	FM	JI	VI	NMI	RI	HA	MA	FM	JI	VI	NMI
k-means++	0.94	0.29	0.30	0.32	0.19	3.88	0.58	0.94	0.29	0.30	0.32	0.19	3.88	0.58
k-harmonic means	**0.95**	**0.31**	**0.32**	**0.34**	**0.20**	**3.83**	**0.59**	**0.95**	**0.31**	**0.32**	**0.34**	**0.20**	**3.83**	**0.59**
GMM-EM	0.50	0.06	0.06	0.24	0.07	4.16	0.26	0.50	0.06	0.06	0.24	0.07	4.16	0.26
Mean Shift	0.65	0.08	0.09	0.25	0.08	4.07	0.36	0.04	0.00	0.00	0.20	0.04	4.64	0.00

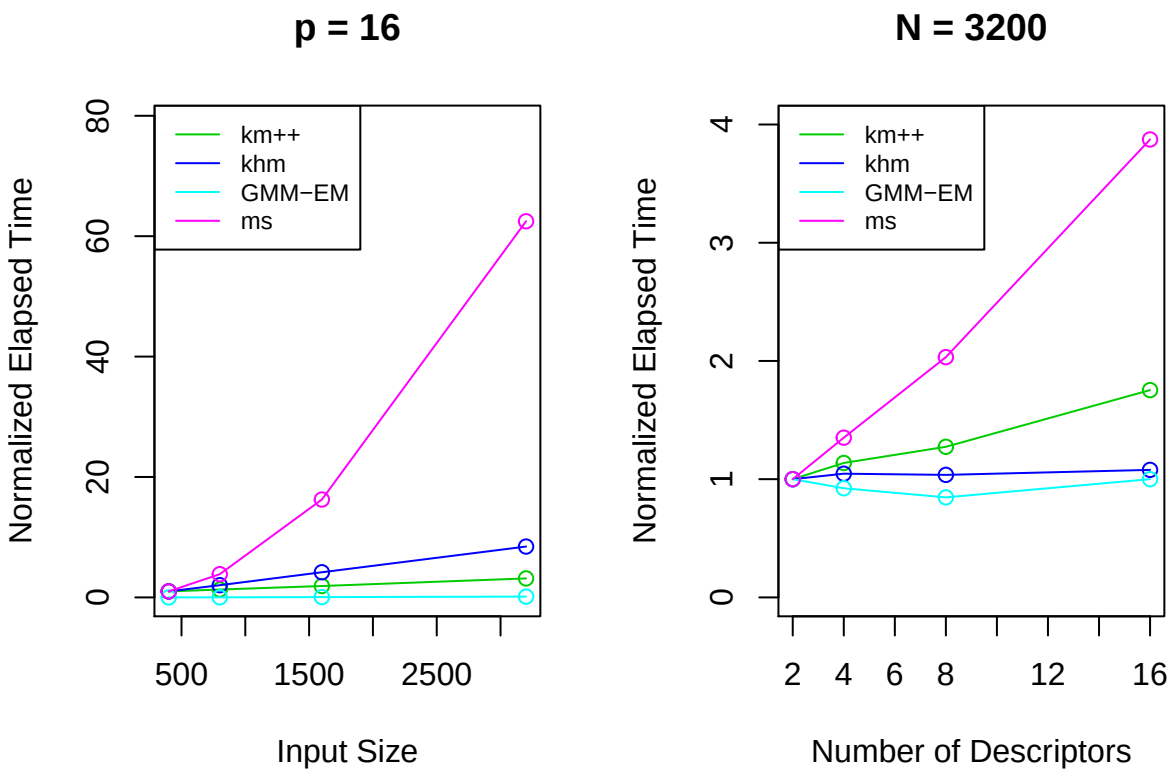
Simulation 5

```
# n = 100, 200, 400, 800, 1600 and k = 2, 4, 8

N <- c(1600) # Number of random samples
p <- c(16, 32, 64, 128, 128*2)
set.seed(123)

times <- list()
for(l in 1:4) {
  times[[l]] <- matrix(ncol = length(p), nrow = length(N))
}

for (i in 1:length(N)) {
  for (j in 1:length(p)) {
    # Target parameters for univariate normal distributions
    mu1 <- -2
    mu2 <- 2
    mu <- c(mu1,mu2) # Mean components
    s <- 1
    c <- sample(1:2,prob=c(0.5,0.5),size=N[i],replace=TRUE)
    mix.mu <- mu[c]

    mydata <- matrix(ncol = p[j], nrow = N[i])

    for(l in 1:N[i]) {
      for(m in 1:p[j]) {
        mydata[l, m] <- rnorm(1, mix.mu[l], sd = s)
      }
    }

    plot(mydata,xlab="X1",ylab="X2", col = c + 2, pch = 19)
    dat <- mydata
    k <- 2
    times[[1]][i, j] <- system.time(kmpp_slow(dat, k))[3]
    times[[2]][i, j] <- system.time(khm(dat, k))[3]
    randPairs <- randomPairs(dat)
    print(dim(dat))
    times[[3]][i, j] <- system.time(Mclust(dat, G = k, modelNames = "EII",
                                           initialization = list(hcPairs = randPairs),
                                           control = emControl(tol = c(0,0), itmax = 100)))[3]
    times[[4]][i, j] <- system.time(ms(dat, thr=0, iter=100, plotms=0))[3]
```

```
    print(times)
  }
}

col1 <- 2

par(mfrow = c(1, 2))

# Use these two plots
plot(1, type = "n", ylab = "Elapsed Time (ms)",
     xlab = "Input Size", xlim = range(N), ylim = c(0, max(times[[4]][, length(p)])),
     main = paste0("p = ", max(p)))
for(i in 1:4) {
  lines(x = N, y = times[[i]][, length(p)], type = "o", col = i+2)
}
legend("topleft", legend=c("k-means++", "k-harmonic means", "GMM-EM", "Mean Shift"),
       col=1:4 + 2, lty = 1)

plot(1, type = "n", ylab = "Elapsed Time (ms)",
     xlab = "Number of Descriptors", xlim = range(p), ylim = c(0, max(times[[4]][length(N), ])),
     main = paste0("N = ", max(N)))
for(i in 1:4) {
  lines(x = p, y = times[[i]][length(N), ], type = "o", col = i+2)
}
legend("topleft", legend=c("k-means++", "k-harmonic means", "GMM-EM", "Mean Shift"),
       col=1:4 + 2, lty = 1)
par(mfrow = c(1, 1))

save.image("run_time_plots.rda")
```

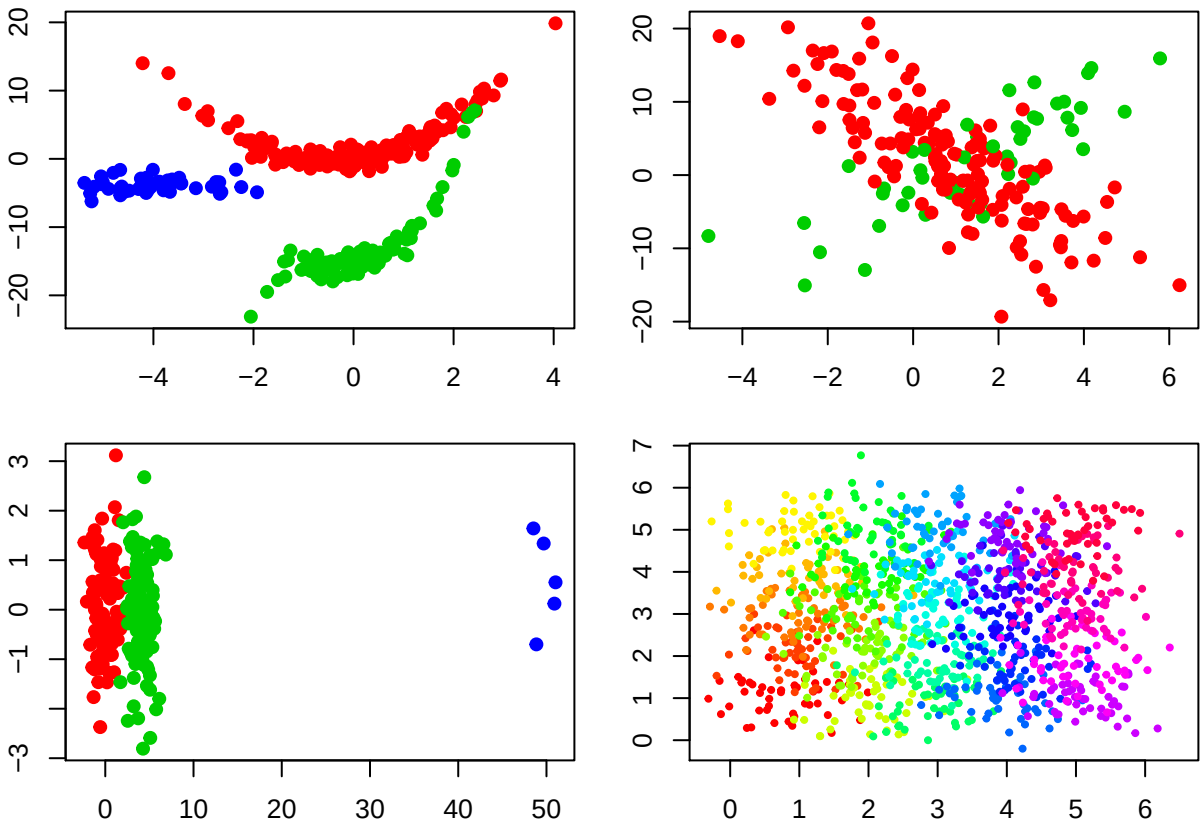
```
load("run_time_plots.rda")
par(mfrow = c(1, 2))
# Use these two plots
plot(1, type = "n", ylab = "Normalized Elapsed Time",
     xlab = "Input Size", xlim = range(N[1:4]), ylim = c(0, max(times[[4]][1:4, 1])),
     main = paste0("p = ", max(p[1:4])))
for(i in 1:4) {
  if(i == 3) {
    lines(x = N[1:4], y = times[[i]][1:4,length(p)], type = "o", col = i+2)
  } else {
    lines(x = N[1:4], y = times[[i]][1:4, length(p)]/times[[i]][1, length(p)], type = "o", col = i+2)
  }
}
legend("topleft", legend=c("km++", "khm", "GMM-EM", "ms"),
       col=1:4 + 2, lty = 1, cex = .75)
L <- 4
plot(1, type = "n", ylab = "Normalized Elapsed Time",
     xlab = "Number of Descriptors", xlim = range(p[1:4]), ylim = c(0, 4),
     main = paste0("N = ", max(N[1:L])))
for(i in 1:4) {
  if(i == 3 ){
    lines(x = p[1:L], y = times[[i]][length(N), 1:L]/times[[i]][length(N), 1], type = "o", col = i+2)
  } else {
    lines(x = p[1:L], y = times[[i]][length(N), 1:L]/times[[i]][length(N), 1], type = "o", col = i+2)
```

```
  }
}
legend("topleft", legend=c("km++", "khm", "GMM-EM", "ms"),
       col=1:4 + 2, lty = 1, cex = .75)
```



Final figure of all data sets

```
par(mfrow=c(2,2), mai = c(.3, 0.3, 0.3, 0.3))
plot(dat1, pch=19, xlab = "", ylab = "x2", col = truth1 + 1)
plot(dat2, pch=19, xlab = "", ylab = "",  col = truth2 + 1)
plot(dat3, pch=19, xlab = "x1", ylab = "x2",  col = truth3 + 1)
plot(dat4, pch=19, xlab = "x1", ylab = "",  col = rainbow(25)[truth4], cex = .5)
```


20
10
0
−10
−20
−4
−2
0
2
4
20
10
0
−10
−20
−4
−2
0
2
4
6
3
2
1
0
−1
−3
0
10
20
30
40
50
7
6
5
4
3
2
1
0
0
1
2
3
4
5
6