

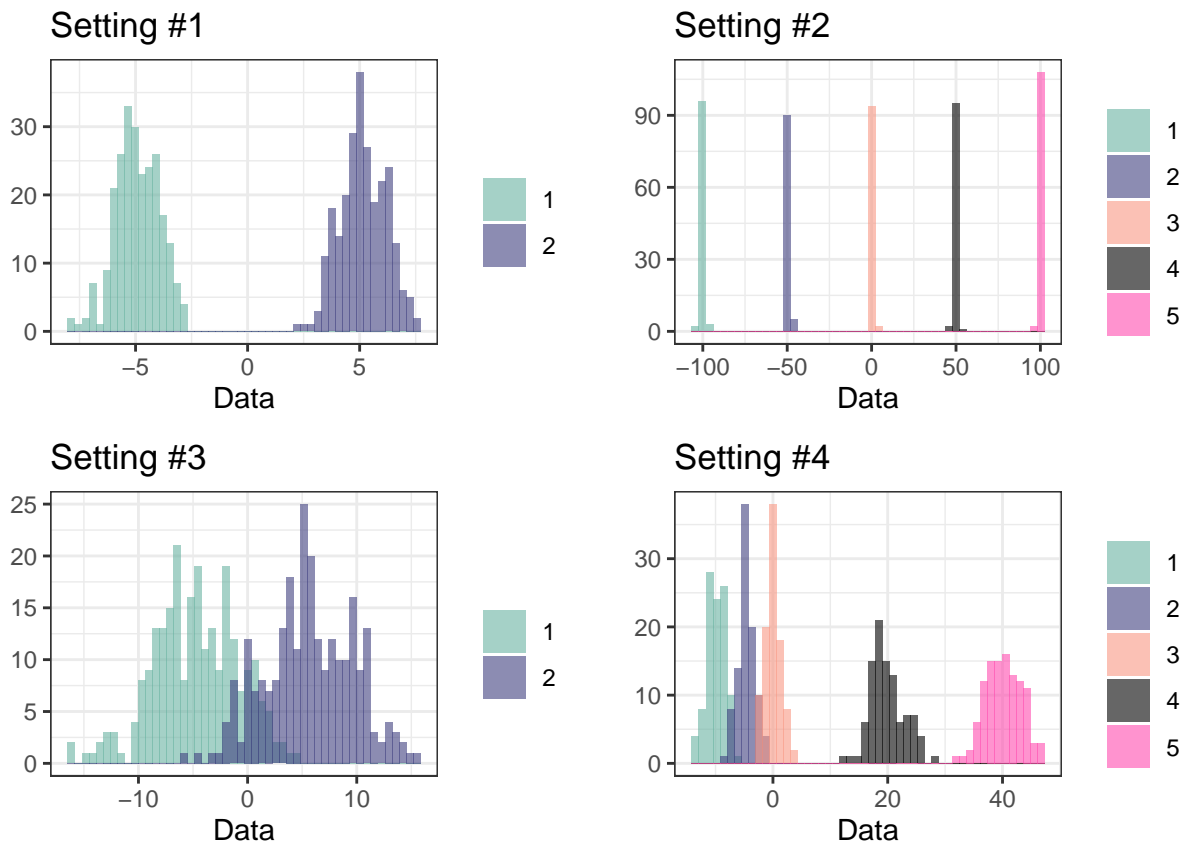
Result - 4/18

Kevin Korsurat

2023-05-01

(0) The simulated data

For the entire analysis in this report, there are 4 settings. The first two settings are the same as those I previously used last time (Result - 3/8). The setting #3 and #4 are the new settings. Below are the histograms showing the simulated data in each setting.



```
n_unique <- function(vec){  
  length(unique(vec))  
}
```

Setting 1

```
### Setting 1
K <- 10
iter <- 1000

ci_init <- sample(1:1, 500, replace = TRUE)
xi_vec <- rep(0.01, K)
mu0_vec <- rep(0, K)
a_sigma_vec <- rep(1, K)
b_sigma_vec <- rep(1, K)
lambda_vec <- rep(1, K)
a_theta <- 1
b_theta <- 1
sm_iter <- 10

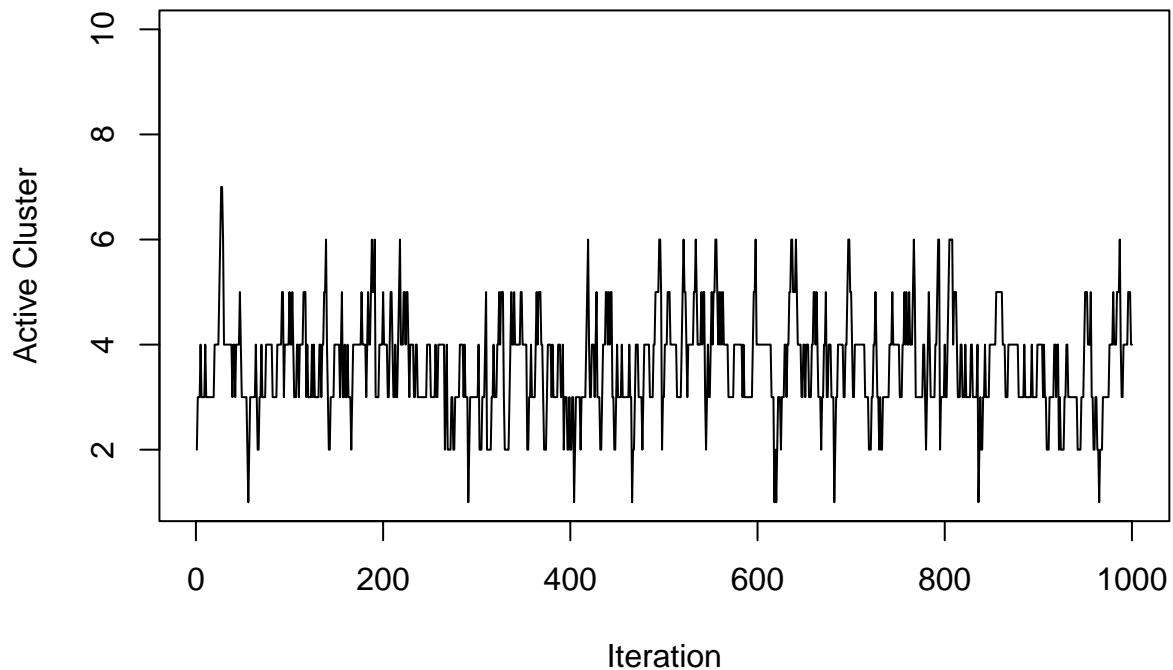
set.seed(seed_val)
start_time <- Sys.time()
result1 <- SFDM_model(iter, K, ci_init, xi_vec, scale(data_sim_1), mu0_vec,
                      a_sigma_vec, b_sigma_vec, lambda_vec, a_theta, b_theta,
                      sm_iter, 250)
Sys.time() - start_time
```

Time difference of 20.01264 secs

```
table(ci_actual_1, salso(result1$iter_assign[-(1:500)], ], maxNClusters = K))
```

```
##
## ci_actual_1    1    2
##              1 246    0
##              2    0 254
```

```
plot(1:iter, apply(result1$iter_assign, 1, n_unique), type = "l",
     ylim = c(1, K), xlab = "Iteration", ylab = "Active Cluster")
```



```
mean(apply(result1$iter_assign, 1, n_unique))
```

```
## [1] 3.596
```

```
result_status <- factor(result1$sm_status)
levels(result_status) <- c("Reject", "Accept")
result_sm <- factor(result1$split_or_merge)
levels(result_sm) <- c("Merge", "Split")
table(result_status, result_sm)
```

```
##           result_sm
## result_status Merge Split
##      Reject   580    76
##      Accept     9   335
```

```
rbind(data.frame(data_sim_1, ci_actual_1,
                 ci_result = as.numeric(salso(result1$iter_assign[-(1:500), ], maxNClusters = K))) %>%
  group_by(ci_actual_1) %>%
  summarise(q = quantile(data_sim_1)) %>%
  rename(cluster = ci_actual_1) %>%
  mutate(type = "Actual", status = paste0("Q", c(0, 1, 2, 3, 4))) %>%
  pivot_wider(names_from = status, values_from = q),
  data.frame(data_sim_1, ci_actual_1,
             ci_result = as.numeric(salso(result1$iter_assign[-(1:500), ], maxNClusters = K))) %>%
```

```

group_by(ci_result) %>%
summarise(q = quantile(data_sim_1)) %>%
rename(cluster = ci_result) %>%
mutate(type = "Model", status = paste0("Q", c(0, 1, 2, 3, 4))) %>%
pivot_wider(names_from = status, values_from = q)

```

```

## 'summarise()' has grouped output by 'ci_actual_1'. You can override using the
## '.groups' argument.
## 'summarise()' has grouped output by 'ci_result'. You can override using the
## '.groups' argument.

```

```

## # A tibble: 4 x 7
## # Groups:   cluster [2]
##   cluster type    Q0    Q1    Q2    Q3    Q4
##   <dbl> <chr> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1         1 Actual -7.72 -5.59 -4.94 -4.21 -2.69
## 2         2 Actual  2.31  4.52  5.13  5.94  7.69
## 3         1 Model -7.72 -5.59 -4.94 -4.21 -2.69
## 4         2 Model  2.31  4.52  5.13  5.94  7.69

```

```

ci_result_1 <- as.numeric(salso(result1$iter_assign[-(1:500)], ], maxNclusters = K))

```

```

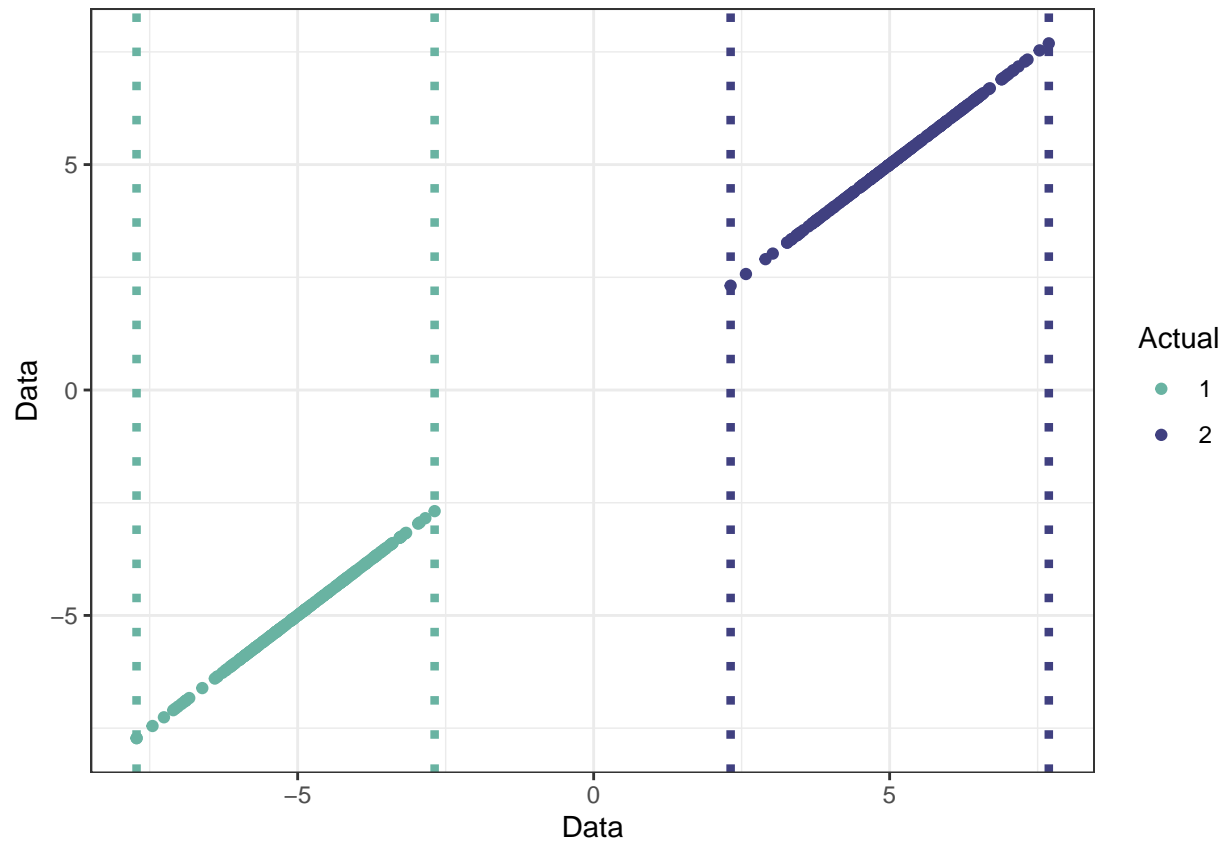
data.frame(data_sim_1, ci_actual_1, ci_result_1) %>%
  ggplot(aes(x = data_sim_1, y = data_sim_1, col = factor(ci_actual_1))) +
  geom_point() +
  theme_bw() +
  scale_color_manual(values=c("#69b3a2", "#404080")) +
  geom_vline(xintercept = quantile(data_sim_1[ci_result_1 == 1], c(0, 1)),
            linetype = "dotted", color = "#69b3a2", size = 1.5) +
  geom_vline(xintercept = quantile(data_sim_1[ci_result_1 == 2], c(0, 1)),
            linetype = "dotted", color = "#404080", size = 1.5) +
  labs(col = "Actual", x = "Data", y = "Data")

```

```

## Warning: Using 'size' aesthetic for lines was deprecated in ggplot2 3.4.0.
## i Please use 'linewidth' instead.

```



Setting 2

```
### Setting 2
K <- 10
iter <- 1000

ci_init <- rep(1, 500)
xi_vec <- rep(0.01, K)
mu0_vec <- rep(0, K)
a_sigma_vec <- rep(100, K)
b_sigma_vec <- rep(1, K)
lambda_vec <- rep(0.01, K)
a_theta <- 1
b_theta <- 4
sm_iter <- 10

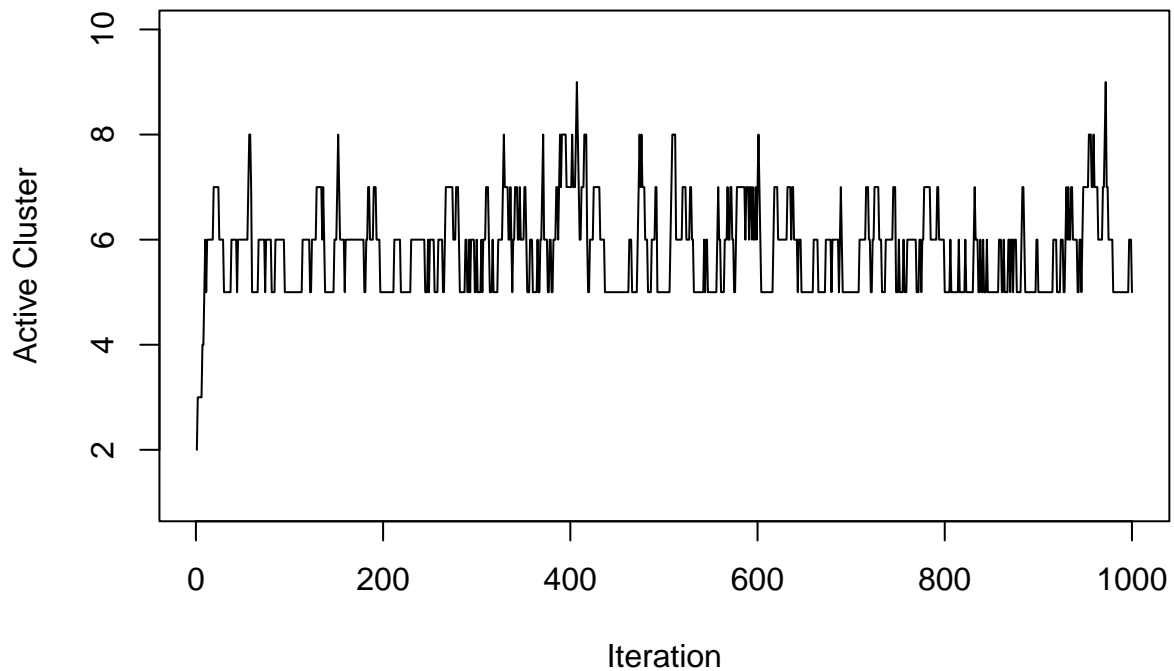
set.seed(seed_val)
start_time <- Sys.time()
result2 <- SFDM_model(iter, K, ci_init, xi_vec, scale(data_sim_2), mu0_vec,
                      a_sigma_vec, b_sigma_vec, lambda_vec, a_theta, b_theta,
                      sm_iter, 250)
Sys.time() - start_time
```

Time difference of 12.5272 secs

```
table(ci_actual_2, salso(result2$iter_assign[-(1:500), ], maxNClusters = K))
```

```
##
## ci_actual_2  1    2    3    4    5
##           1  0 101    0    0    0
##           2  0   0    0  95    0
##           3  0   0    0   0  96
##           4  0   0   98   0   0
##           5 110   0   0   0   0
```

```
plot(1:iter, apply(result2$iter_assign, 1, n_unique), type = "l",
     ylim = c(1, K), xlab = "Iteration", ylab = "Active Cluster")
```



```
mean(apply(result2$iter_assign, 1, n_unique))
```

```
## [1] 5.802
```

```
result_status <- factor(result2$sm_status)
levels(result_status) <- c("Reject", "Accept")
result_sm <- factor(result2$split_or_merge)
levels(result_sm) <- c("Merge", "Split")
table(result_status, result_sm)
```

```
##               result_sm
## result_status Merge Split
##      Reject    801    23
##      Accept     0   176
```

Setting 3

```
### Setting 3
K <- 10
iter <- 1000
ci_init <- rep(1:1, 500)
xi_vec <- rep(0.01, K)
mu0_vec <- rep(0, K)
a_sigma_vec <- rep(100, K)
b_sigma_vec <- rep(10, K)
lambda_vec <- rep(10, K)
a_theta <- 1
b_theta <- 1
sm_iter <- 10

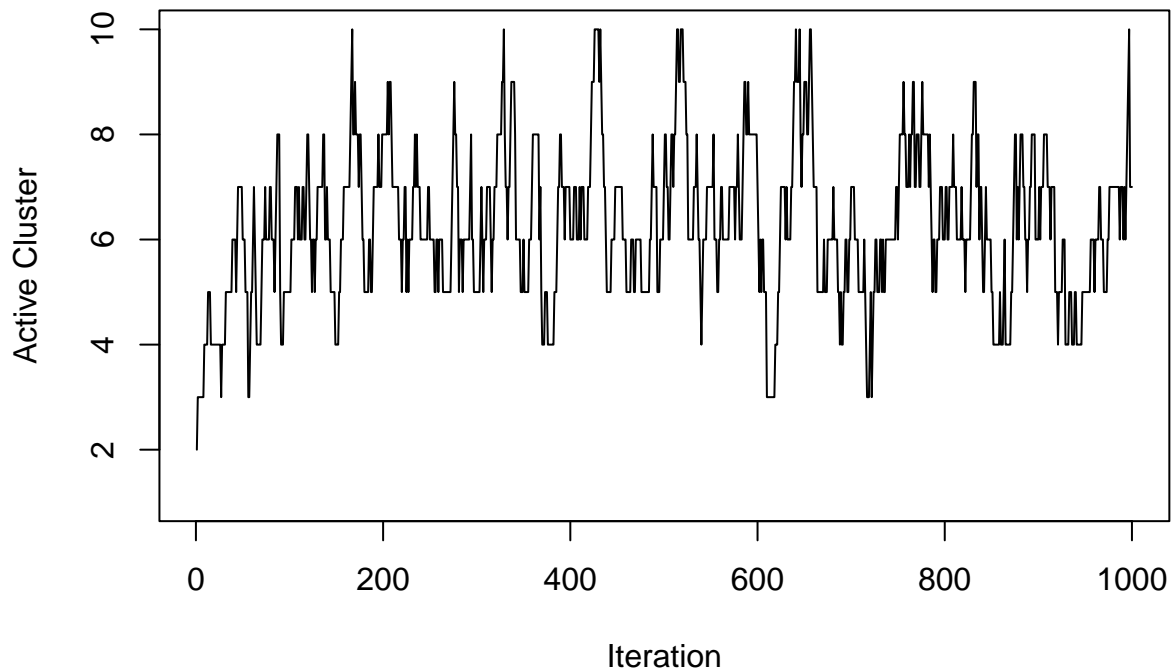
set.seed(seed_val)
start_time <- Sys.time()
result3 <- SFDM_model(iter, K, ci_init, xi_vec, scale(data_sim_3), mu0_vec,
                      a_sigma_vec, b_sigma_vec, lambda_vec, a_theta, b_theta,
                      sm_iter, 250)
Sys.time() - start_time
```

```
## Time difference of 17.82747 secs
```

```
table(ci_actual_3, salso(result3$iter_assign[-(1:500), ], maxNClusters = K))
```

```
##
## ci_actual_3    1    2
##           1 235  11
##           2  46 208
```

```
plot(1:iter, apply(result3$iter_assign, 1, n_unique), type = "l",
     ylim = c(1, K), xlab = "Iteration", ylab = "Active Cluster")
```



```
mean(apply(result3$iter_assign, 1, n_unique))
```

```
## [1] 6.244
```

```
result_status <- factor(result3$sm_status)
levels(result_status) <- c("Reject", "Accept")
result_sm <- factor(result3$split_or_merge)
levels(result_sm) <- c("Merge", "Split")
table(result_status, result_sm)
```

```
##           result_sm
## result_status Merge Split
##      Reject   696    13
##      Accept     0   291
```

```
rbind(data.frame(data_sim_3, ci_actual_3,
                 ci_result = as.numeric(salso(result3$iter_assign[-(1:500), ], maxNClusters = K))) %>%
  group_by(ci_actual_3) %>%
  summarise(q = quantile(data_sim_3)) %>%
  rename(cluster = ci_actual_3) %>%
  mutate(type = "Actual", status = paste0("Q", c(0, 1, 2, 3, 4))) %>%
  pivot_wider(names_from = status, values_from = q),
  data.frame(data_sim_3, ci_actual_3,
             ci_result = as.numeric(salso(result3$iter_assign[-(1:500), ], maxNClusters = K))) %>%
```



```

group_by(ci_result) %>%
  summarise(q = quantile(data_sim_3)) %>%
  rename(cluster = ci_result) %>%
  mutate(type = "Model", status = paste0("Q", c(0, 1, 2, 3, 4))) %>%
  pivot_wider(names_from = status, values_from = q)

```

'summarise()' has grouped output by 'ci_actual_3'. You can override using the
 ## '.groups' argument.
 ## 'summarise()' has grouped output by 'ci_result'. You can override using the
 ## '.groups' argument.

```

## # A tibble: 4 x 7
## # Groups:   cluster [2]
##   cluster type    Q0    Q1    Q2    Q3    Q4
##   <dbl> <chr>   <dbl> <dbl> <dbl> <dbl> <dbl>
## 1      1 Actual -15.9 -7.38 -4.78 -1.84  4.25
## 2      2 Actual  -5.74  3.08  5.52  8.74 15.8
## 3      1 Model -15.9 -6.99 -4.27 -1.23  1.66
## 4      2 Model   1.81  4.21  6.00  9.31 15.8

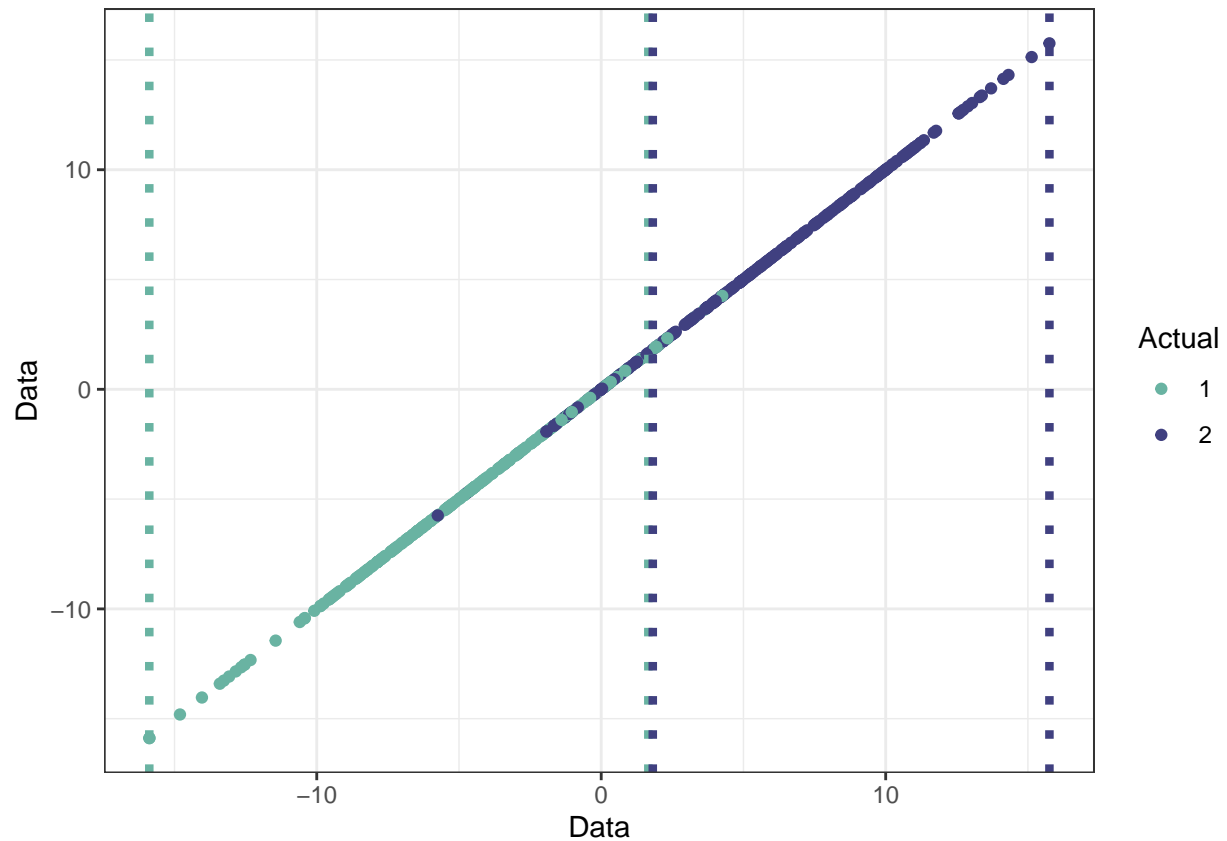
```

```
ci_result_3 <- as.numeric(salso(result3$iter_assign[-(1:500), ], maxNClusters = K))
```

```

data.frame(data_sim_3, ci_actual_3, ci_result_3) %>%
  ggplot(aes(x = data_sim_3, y = data_sim_3, col = factor(ci_actual_3))) +
  geom_point() +
  theme_bw() +
  scale_color_manual(values=c("#69b3a2", "#404080")) +
  geom_vline(xintercept = quantile(data_sim_3[ci_result_3 == 1], c(0, 1)),
    linetype = "dotted", color = "#69b3a2", size = 1.5) +
  geom_vline(xintercept = quantile(data_sim_3[ci_result_3 == 2], c(0, 1)),
    linetype = "dotted", color = "#404080", size = 1.5) +
  ## geom_vline(xintercept = quantile(data_sim_3[ci_result_3 == 3], c(0, 1)),
  ##           linetype = "dotted", color = "red", size = 1.5) +
  labs(col = "Actual", x = "Data", y = "Data")

```



Setting 4

```
### Setting 4
K <- 10
iter <- 1000

ci_init <- rep(1, 500)
xi_vec <- rep(0.01, K)
mu0_vec <- rep(0, K)
a_sigma_vec <- rep(100, K)
b_sigma_vec <- rep(1, K)
lambda_vec <- rep(1, K)
a_theta <- 1
b_theta <- 1
sm_iter <- 10

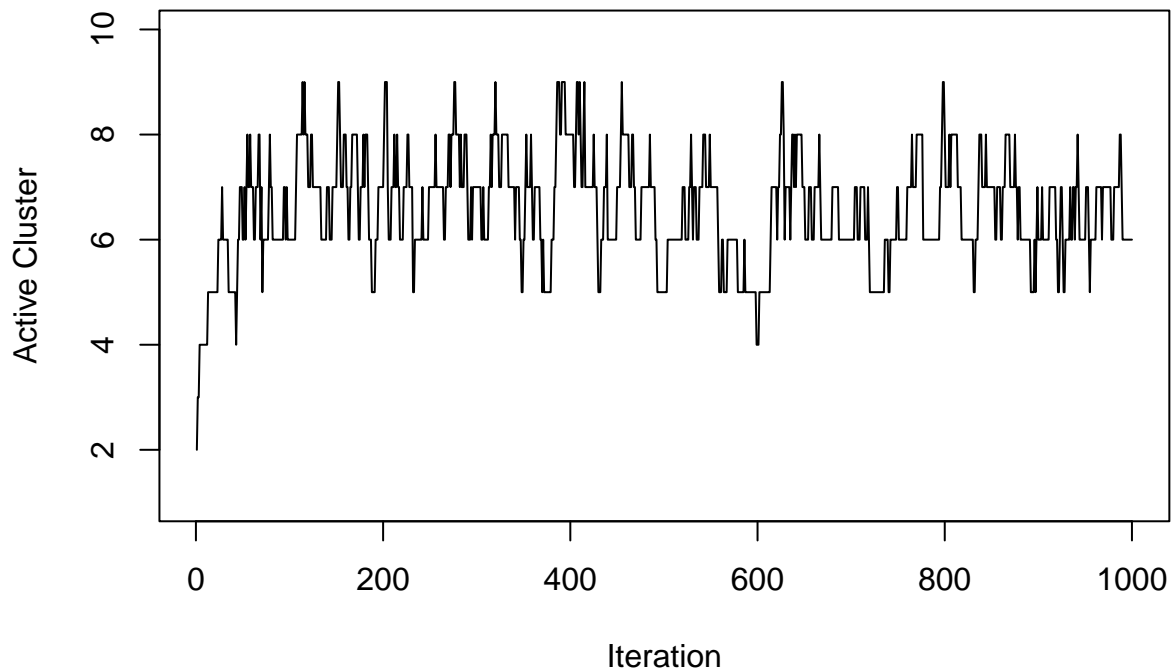
set.seed(seed_val)
start_time <- Sys.time()
result4 <- SFDM_model(iter, K, ci_init, xi_vec, scale(data_sim_4), mu0_vec,
                      a_sigma_vec, b_sigma_vec, lambda_vec, a_theta, b_theta,
                      sm_iter, 250)
Sys.time() - start_time
```

Time difference of 13.33995 secs

```
table(ci_actual_4, salso(result4$iter_assign[-(1:500), ], maxNClusters = K))
```

```
##
## ci_actual_4   1    2    3    4
##           1    0 101    0    0
##           2    0  69    0  26
##           3    0   0    0  96
##           4    0   0   98   0
##           5  110   0   0   0
```

```
plot(1:iter, apply(result4$iter_assign, 1, n_unique), type = "l",
     ylim = c(1, K), xlab = "Iteration", ylab = "Active Cluster")
```



```
mean(apply(result4$iter_assign, 1, n_unique))
```

```
## [1] 6.561
```

```
result_status <- factor(result4$sm_status)
levels(result_status) <- c("Reject", "Accept")
result_sm <- factor(result4$split_or_merge)
levels(result_sm) <- c("Merge", "Split")
table(result_status, result_sm)
```

```
##           result_sm
## result_status Merge Split
##      Reject    797    14
##      Accept     0    189
```

```
rbind(data.frame(data_sim_4, ci_actual_4,
                 ci_result = as.numeric(salso(result4$iter_assign[-(1:500)], ], maxNClusters = K))) %>%
  group_by(ci_actual_4) %>%
  summarise(q = quantile(data_sim_4)) %>%
  rename(cluster = ci_actual_4) %>%
  mutate(type = "Actual", status = paste0("Q", c(0, 1, 2, 3, 4))) %>%
  pivot_wider(names_from = status, values_from = q),
data.frame(data_sim_4, ci_actual_4,
           ci_result = as.numeric(salso(result4$iter_assign[-(1:500)], ], maxNClusters = K))) %>%
  group_by(ci_result) %>%
  summarise(q = quantile(data_sim_4)) %>%
  rename(cluster = ci_result) %>%
  mutate(type = "Model", status = paste0("Q", c(0, 1, 2, 3, 4))) %>%
  pivot_wider(names_from = status, values_from = q))
```

```
## 'summarise()' has grouped output by 'ci_actual_4'. You can override using the
## '.groups' argument.
## 'summarise()' has grouped output by 'ci_result'. You can override using the
## '.groups' argument.
```

```
## # A tibble: 9 x 7
## # Groups:   cluster [5]
##   cluster type      Q0      Q1      Q2      Q3      Q4
##   <dbl> <chr>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1     1 Actual -13.6 -11.1  -9.97  -8.59  -6.35
## 2     2 Actual  -8.15 -5.49  -4.77  -3.80  -1.09
## 3     3 Actual  -2.94 -0.972 -0.0204 0.829  3.59
## 4     4 Actual  12.6  17.9  19.4   21.8   28.1
## 5     5 Actual  31.9  38.3  40.1   42.4   46.5
## 6     1 Model  31.9  38.3  40.1   42.4   46.5
## 7     2 Model -13.6 -10.3  -8.28  -5.39  -4.05
## 8     3 Model  12.6  17.9  19.4   21.8   28.1
## 9     4 Model  -3.95 -1.94  -0.549  0.524  3.59
```

```
ci_result_4 <- as.numeric(salso(result4$iter_assign[-(1:500)], ], maxNClusters = K))

data.frame(data_sim_4, ci_actual_4, ci_result_4) %>%
  ggplot(aes(x = data_sim_4, y = data_sim_4, col = factor(ci_actual_4))) +
  geom_point() +
  theme_bw() +
  ## scale_color_manual(values=c("#69b3a2", "#404080")) +
  geom_vline(xintercept = quantile(data_sim_4[ci_result_4 == 1], c(0, 1)),
             linetype = "dotted", color = "#69b3a2", size = 1.5) +
  geom_vline(xintercept = quantile(data_sim_4[ci_result_4 == 2], c(0, 1)),
             linetype = "dotted", color = "#404080", size = 1.5) +
  geom_vline(xintercept = quantile(data_sim_4[ci_result_4 == 3], c(0, 1)),
             linetype = "dotted", color = "red", size = 1.5) +
  geom_vline(xintercept = quantile(data_sim_4[ci_result_4 == 4], c(0, 1)),
```

```
linetype = "dotted", color = "orange", size = 1.5) +
labs(col = "Actual", x = "Data", y = "Data")
```

