Simulation Study - AntMAN and DP

Kevin Korsurat

2023-06-20

```r
### Function: Simulating the data based on the scenario
f_data_sim <- function(sim_seed, scenario_index){

  ### place for storing result.
  actual_clus <- NULL
  dat <- NULL

  set.seed(sim_seed)

  if(! scenario_index %in% 1:4){
    warning("invalid scenario. we have only 4 scenarios")
  } else {
    if(scenario_index == 1){
      actual_clus <- sample(1:2, 500, replace = TRUE)
      dat <- rnorm(500, c(-5, 5)[actual_clus])
    } else if(scenario_index == 2){
      actual_clus <- sample(1:5, 500, replace = TRUE)
      dat <- rnorm(500, (c(0, 7.5, 15, 25, 35))[actual_clus])
    } else if(scenario_index == 3){
      actual_clus <- sample(1:2, 500, replace = TRUE)
      dat <- rnorm(500, c(-5, 5)[actual_clus], 3)
    } else {
      actual_clus <- sample(1:5, 500, replace = TRUE)
      dat <- rnorm(500, (c(0, 7.5, 15, 25, 35)[actual_clus])/2, 1)
    }
  }

  ### return the simulated data
  result <- data.frame(actual_clus, dat)
  return(result)
}
```

Sparse Finite Continuous Mixture Model (SFCMM)

We still have three steps for SFCMM, which is similar to our model. (SFDMM)

Reallocation step

Since all clusters are already active for SFCMM, which is different from our model, each observation can be reallocated to any possible clusters, which is similar to the finite mixture model. (For our model, the observations can be reallocated to the already active cluster only.)

Split-Merge step

Since all clusters are already active, we need to set all inclusion indicators to 1. Therefore, we can remove the a_θ and b_θ from the acceptance ratio.

Parameters update

Instead of updating the parameters $\left(\mu_k, \sigma_k^2, \alpha_k\right)$ for the active clusters as in our model, SFCMM will update the parameters for all clusters even though that cluster is empty.

SPCMM and SPDMM (our model)

For both models, I will use the same set of hyperparameters. I will run the result on all scenarios for both raw and scaled data.

- $\text{K}_{\text{max}} = 10$
- $\sigma_0^2 = 100$
- $a_\sigma = b_\sigma = 0.1$
- $\xi = 1$
- $a_\theta = b_\theta = 1$
- the number of launch step is 10

Here is the result for the raw data.

```
### Raw data
for(i in 1:4){
  dat_sim <- f_data_sim(74531, i)
  dat_y <- dat_sim$dat

  print(paste0("============== Scenario ", i, " (Raw Data) =============="))

  ### SFDMM
  model <- SFDMM_model(iter = 10000, K_max = 10, init_assign = rep(0, 500),
                       y = dat_y, a0 = 0.01, b0 = 0.01, mu0 = 0, s20 = 100,
                       xi0 = 1, a_theta = 1, b_theta = 1, launch_iter = 10,
                       print_iter = 10001)
  table("SFDMM" = salso(model$iter_assign[-(1:5000), ]),
        "Actual" = dat_sim$actual_clus) %>% print()

  ### SPCMM
  model <- SFCMM_model(iter = 10000, K_max = 10, init_assign = rep(0, 500),
                       y = dat_y, a0 = 0.01, b0 = 0.01, mu0 = 0, s20 = 100,
                       xi0 = 1, launch_iter = 10, print_iter = 10001)
  table("SFCMM" = salso(model$iter_assign[-(1:5000), ]),
        "Actual" = dat_sim$actual_clus) %>% print()

}
```

```
## [1] "============== Scenario 1 (Raw Data) =============="
##       Actual
## SFDMM   1   2
##     1 236   0
##     2   0 264
##       Actual
## SFCMM   1   2
##     1 236   0
##     2   0 264
## [1] "============== Scenario 2 (Raw Data) =============="
##       Actual
## SFDMM   1   2   3   4   5
##     1 103   0   0   0   0
##     2   0   0  90   0   0
##     3   0 108   0   0   0
##     4   0   0   0  98   0
##     5   0   0   0   0 101
##       Actual
## SFCMM   1   2   3   4   5
##     1 103   0   0   0   0
##     2   0   0  90   0   0
##     3   0 108   0   0   0
##     4   0   0   0  98   0
##     5   0   0   0   0 101
## [1] "============== Scenario 3 (Raw Data) =============="
##       Actual
## SFDMM   1   2
##     1   8 237
##     2 228  27
##       Actual
## SFCMM   1   2
##     1  10 241
##     2 226  23
## [1] "============== Scenario 4 (Raw Data) =============="
##       Actual
## SFDMM   1   2   3   4   5
##     1 103   6   0   0   0
##     2   0   9  87   3   0
##     3   0   0   1  95   0
##     4   0   0   0   0 101
##     5   0  93   2   0   0
##       Actual
## SFCMM   1   2   3   4   5
##     1 103   6   0   0   0
##     2   0   9  87   3   0
##     3   0   0   1  95   0
##     4   0   0   0   0 101
##     5   0  93   2   0   0
```

Here is the result for the scaled data.

```r
### Raw data
for(i in 1:4){
  dat_sim <- f_data_sim(55430, i)
  dat_y <- as.numeric(scale(dat_sim$dat))

  print(paste0("============== Scenario ", i, " (Scaled Data) =============="))
  ### SFDMM
  model <- SFDMM_model(iter = 10000, K_max = 10, init_assign = rep(0, 500),
                       y = dat_y, a0 = 0.01, b0 = 0.01, mu0 = 0, s20 = 100,
                       xi0 = 1, a_theta = 1, b_theta = 1, launch_iter = 10,
                       print_iter = 10001)
  table("SFDMM" = salso(model$iter_assign[-(1:5000), ]),
        "Actual" = dat_sim$actual_clus) %>% print()

  ### SPCMM
  model <- SFCMM_model(iter = 10000, K_max = 10, init_assign = rep(0, 500),
                       y = dat_y, a0 = 0.01, b0 = 0.01, mu0 = 0, s20 = 100,
                       xi0 = 1, launch_iter = 10, print_iter = 10001)
  table("SFCMM" = salso(model$iter_assign[-(1:5000), ]),
        "Actual" = dat_sim$actual_clus) %>% print()

}
```

```
## [1] "============== Scenario 1 (Scaled Data) =============="
##      Actual
## SFDMM   1   2
##     1   0 254
##     2 246   0
##      Actual
## SFCMM   1   2
##     1   0 254
##     2 246   0
## [1] "============== Scenario 2 (Scaled Data) =============="
##      Actual
## SFDMM   1   2   3   4   5
##     1  97   0   0   0   0
##     2   0   0   0 109   0
##     3   0   0   0   0  93
##     4   0   0 105   0   0
##     5   0  96   0   0   0
##      Actual
## SFCMM   1   2   3   4   5
##     1  97   0   0   0   0
##     2   0   0   0 109   0
##     3   0   0   0   0  93
##     4   0   0 105   0   0
##     5   0  96   0   0   0
## [1] "============== Scenario 3 (Scaled Data) =============="
##      Actual
## SFDMM   1   2
##     1  16 243
##     2 230  11
##      Actual
## SFCMM   1   2
##     1  14 243
##     2 232  11
## [1] "============== Scenario 4 (Scaled Data) =============="
##      Actual
## SFDMM   1   2   3   4   5
##     1  95   7   0   0   0
##     2   0   0   0 109   1
##     3   0   0   0   0  92
##     4   0   4 100   0   0
##     5   2  85   5   0   0
##       Actual
## SFCMM   1   2   3   4   5
##     1  95   7   0   0   0
##     2   0   0   0 109   1
##     3   0   0   0   0  92
##     4   0   4 100   0   0
##     5   2  85   5   0   0
```